Data privacy: an overview

Vicenç Torra

December, 2019

Hamilton Institute, Maynooth University, Ireland

Overview

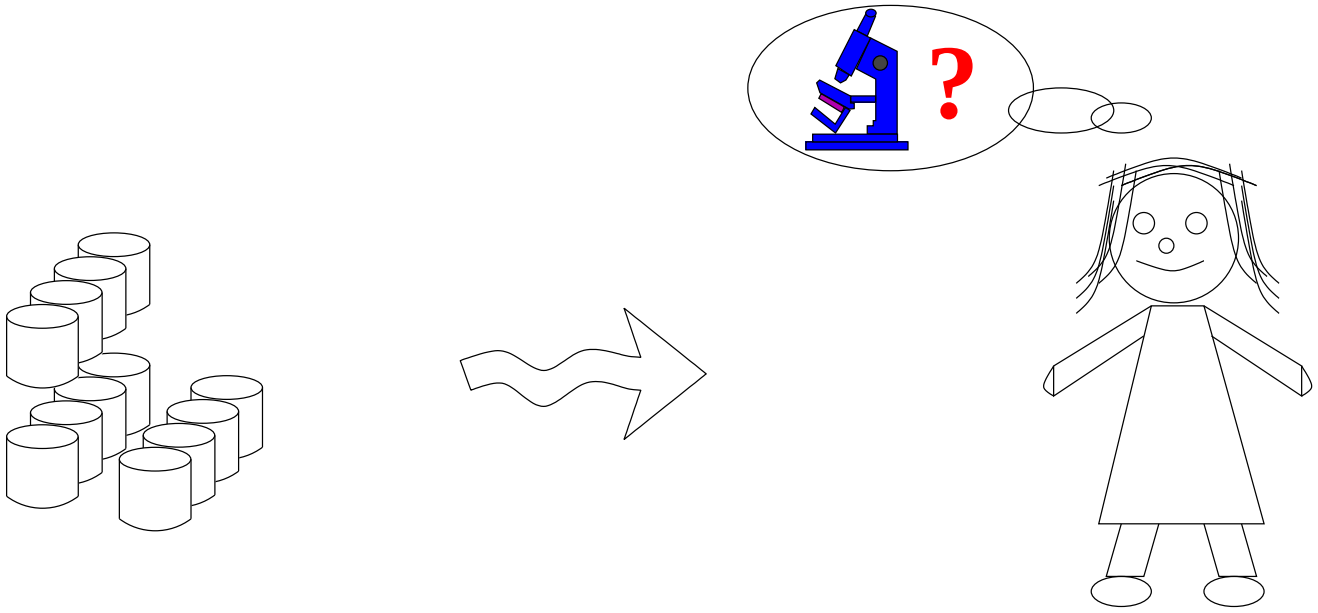
- What is data privacy?
- Why is it necessary and why it is challenging/difficult?
- Some definitions
- Privacy models
- Privacy methods

Outline

Motivation

Introduction

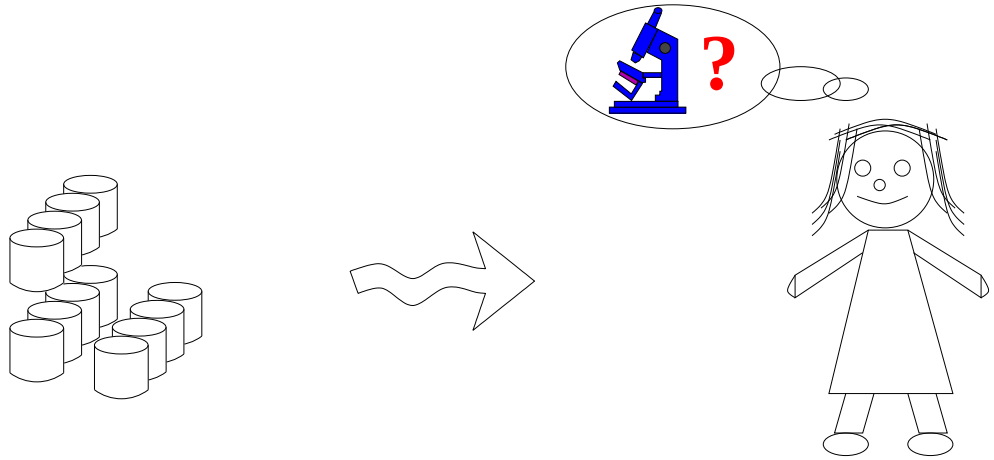
- Data privacy: core
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.



E.g., you are authorized to compute the average stay in a hospital, but you are not authorized to see the length of stay of your neighbor.

Introduction

- Data privacy: core
  - (Someone $\Rightarrow$ A third party) accesses data for an authorized analysis, but access and the results should avoid disclosure.
    $\Rightarrow$ The third party can be external to the company or **internal** with restricted access. E.g., admissions in hospital with no access to diagnosis, technician in a bank with no access to credit card records.



E.g., you are authorized to compute the average stay in a hospital, but you are not authorized to see the length of stay of your neighbor.

Introduction

- Problems/difficulties?

Introduction

- Problems/difficulties?
  - Sensitive information

Introduction

- Problems/difficulties?
  - Sensitive information
  - the data

Introduction

- Problems/difficulties?
  - Sensitive information
  - the data

    access to the original data

Introduction

- Problems/difficulties?
  - Sensitive information
  - the data
    access to the original data
  - the outcome/aggregate

Introduction

- Problems/difficulties?
  - Sensitive information
  - the data
    access to the original data
  - the outcome/aggregate
    the solution is leakage of information

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$
    is this ok ?

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",

    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$

    is this ok ?

    NO!!:

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$
    is this ok ?
    NO!!:
  - E.g., there is only one student of anthropology living in Enfield.
    (Enfield, Anthropology, Yes)

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$
    is this ok ?
    NO!!:
  - E.g., there is only one student of anthropology living in Enfield.
    (Enfield, Anthropology, Yes)

  $\Rightarrow$ 1. We learn that our friend is in the database

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$
    is this ok ?
    NO!!:
  - E.g., there is only one student of anthropology living in Enfield.
    (Enfield, Anthropology, Yes)

$\Rightarrow$ 1. We learn that our friend is in the database

$\Rightarrow$ 2. We learn that our friend is sick !!

Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?

Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Example[1]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300

[1] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Example[1]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300
  - Mean income is not "personal data", is this ok ?

[1] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Example[1]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300
  - Mean income is not "personal data", is this ok ?
    NO!!:

[1] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

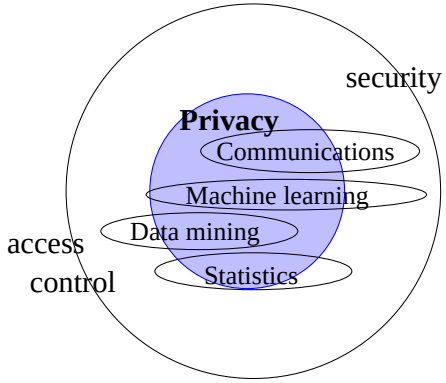
Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Example[1]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300
  - Mean income is not "personal data", is this ok ?
    NO!!:
  - Adding Ms. Rich's salary 100,000 Eur/month: mean = 12090,90 !
    (a extremely high salary changes the mean significantly)
    $\Rightarrow$ We infer Ms. Rich from Town was attending the unit

[1] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

Introduction

- A personal view of core and boundaries of data privacy: core
  - data uses / rellevant techniques
    - ⋆ Data to be used for data analysis
      - ⇒ statistics, machine learning, data mining
      - ⇒ compute indices, find patterns, build models
    - ⋆ Data is transmitted
      - ⇒ communications
      - ⇒ protecting sender identity



- Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.

Introduction

- A personal view of core and boundaries of data privacy: boundaries
  - Database in a computer or in a removable device

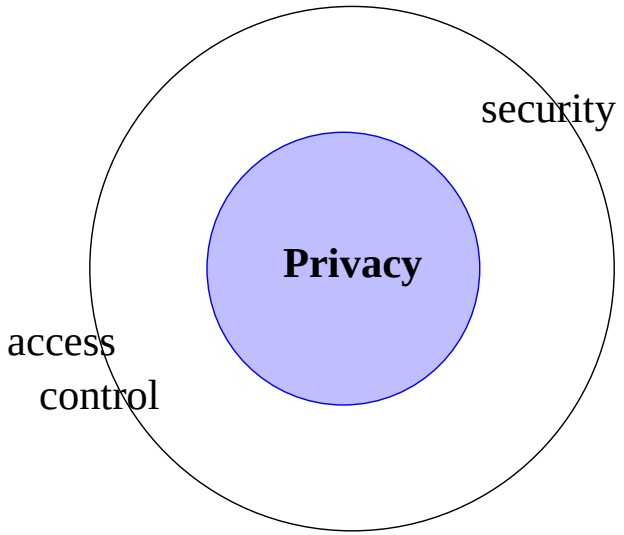
    $\Rightarrow$ access control to avoid unauthorized access

    $\Longrightarrow$ Access to address (admissions), Access to blood test (admissions?)
  - Data is transmitted

    $\Rightarrow$ security technology to avoid unauthorized access

    $\Longrightarrow$ Data from blood glucose meter sent to hospital. Network sniffers

    Transmission is sensitive: Near miss/hit report to car manufacturers



Motivation

Motivation I

- Legislation
  - Privacy a fundamental right. (Ch. 1.1)
    - ⋆ Universal Declaration of Human Rights (UN). European Convention on Human Rights (Council of Europe). General Data Protection Regulation - GDPR (EU). National regulations.
  - Enforcement (GDPR)
    - ⋆ Obligations with respect to data processing
    - ⋆ Requirement to report personal data breaches
    - ⋆ Grant individual rights (to be informed, to access, to rectification, to erasure, ...)

Motivation

Motivation II

- Companies own interest.
  - Competitors can take advantage of information.
  - Privacy-friendly

    (e.g. https://secuso.aifb.kit.edu/english/105.php)

    $\Rightarrow$ Socially responsible company
- Avoiding privacy breaches.
  - Several well known cases.

    $\Rightarrow$ Corporate image

Motivation

- Privacy and society
  - Not only a computer science/technical problem
    - ⋆ Social roots of privacy
    - ⋆ Multidisciplinary problem
  - Social, legal, philosophical questions

Motivation

- Privacy and society
  - Not only a computer science/technical problem
    - ⋆ Social roots of privacy
    - ⋆ Multidisciplinary problem
  - Social, legal, philosophical questions
  - Culturally relative?

    I.e., the importance of privacy is the same among all people ?

Motivation

- Privacy and society
  - Not only a computer science/technical problem
    - ⋆ Social roots of privacy
    - ⋆ Multidisciplinary problem
  - Social, legal, philosophical questions
  - Culturally relative?

    I.e., the importance of privacy is the same among all people ?
  - Are there aspects of life which are inherently private or just conventionally so?

Motivation

- Privacy and society
  - Not only a computer science/technical problem
    - ⋆ Social roots of privacy
    - ⋆ Multidisciplinary problem
  - Social, legal, philosophical questions
  - Culturally relative?
    I.e., the importance of privacy is the same among all people ?
  - Are there aspects of life which are inherently private or just conventionally so?

- This has implications: e.g. tension between privacy and security.
  Different perspectives lead
  - to different solutions and *privacy levels*
  - and to different variables to protect.

Motivation

- Privacy and society. Is this a new problem? Yes and not

Motivation

- Privacy and society. Is this a new problem? Yes and not
  - No side. See the following:

    Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops." (...)
    Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery (...) To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle.
    (S. D. Warren and L. D. Brandeis, 1890)

Motivation

- Privacy and society. Is this a new problem? Yes and not
  - No side. See the following:

    Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops." (...)
    Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery (...) To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle.
    (S. D. Warren and L. D. Brandeis, 1890)
  - Yes side. Big data, storage, mobile, surveillance/CCTV, RFID, IoT
    $\Rightarrow$ pervasive tracking

Motivation

- Technical solutions for data privacy (details later)
  - Statistical disclosure control (SDC)
  - Privacy enhancing technologies (PET)
  - Privacy preserving data mining (PPDM)
- Socio-technical aspects
  - Technical solutions are not enough
  - Implementation/management of solutions for achieving data privacy need to have a holistic perspective of information systems
  - E.g., employees and customers: how technology is applied

Motivation

- Technical solutions for data privacy (details later)
  - Statistical disclosure control (SDC)
  - Privacy enhancing technologies (PET)
  - Privacy preserving data mining (PPDM)
- Socio-technical aspects
  - Technical solutions are not enough
  - Implementation/management of solutions for achieving data privacy need to have a holistic perspective of information systems
  - E.g., employees and customers: how technology is applied
    $\Rightarrow$ we can implement access control and data privacy, but if a printed copy of a *confidential* transaction is left in the printer . . . , or captured with a camera . . .

Motivation

- Technical solutions for data privacy from
  - Statistical disclosure control (SDC)
    - ⋆ Protection for statistical surveys and census
    - ⋆ National statistical offices
    - ⋆ (Dalenius, 1977)

Motivation

- Technical solutions for data privacy from
  - Statistical disclosure control (SDC)
    - ⋆ Protection for statistical surveys and census
    - ⋆ National statistical offices
    - ⋆ (Dalenius, 1977)
  - Privacy enhancing technologies (PET)
    - ⋆ Protection for communications / data transmission
    - ⋆ E.g., anonymous communications (Chaum 1981)

Motivation

- Technical solutions for data privacy from
  - Statistical disclosure control (SDC)
    - ⋆ Protection for statistical surveys and census
    - ⋆ National statistical offices
    - ⋆ (Dalenius, 1977)
  - Privacy enhancing technologies (PET)
    - ⋆ Protection for communications / data transmission
    - ⋆ E.g., anonymous communications (Chaum 1981)
  - Privacy preserving data mining (PPDM)
    - ⋆ Data mining for databases
    - ⋆ Data from banks, hospitals, and economic transactions (late 1990s)

Difficulties

Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[2]

  Names, Age, Sex, Occupation, Place of birth, Last place of residence, Yes/No, condition (healthy?)

LIST OF PASSENGERS.

REPORT AND LIST OF PASSENGERS taken on board the S.S. Hooper at London.

NAMES	AGE	SEX	OCCUPATION	PLACE OF BIRTH	Last Place of Residence	CONDITION
[illegible]	11	Male	Gentleman	Aberdeen Eng	Chelmsford Eng	Healthy
[illegible]	24		Lawyer's Clerk	Madras India	Saltburn Eng	
[illegible]	21		Tailor	London	London	
[illegible]	26		Surgeon	London	London	
[illegible]	28		Gentleman	Boston U.S.	Boston U.S.	
[illegible]	30			Scotland	Boston U.S.	
[illegible]	16			London (?)	London	
[illegible]	44			Dover Eng.	Dover Eng.	
[illegible]	28			Aberdeen Scotland	Boston U.S.	
[illegible]	28		Cattle Dealer	Boston U.S.	Boston U.S.	

[2] https://www.sec.state.ma.us/arc/arcgen/genidx.htm

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - ⋆ 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - ⋆ 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

- A few variables suffice for identifying someone. They are not "personal"

Difficulties

- Difficulties: highly identifiable data
  - An only record *(25 years old, town)*
    all other records with *(age $>$ 35, town)*

- A few variables suffice for identifying someone. They are not "personal"

Difficulties

- Difficulties: highly identifiable data
  - Data from mobile devices:
    $\Rightarrow$ two positions can make you unique (home and working place)

Difficulties

- Difficulties: highly identifiable data
  - Data from mobile devices:
    $\Rightarrow$ two positions can make you unique (home and working place)

- A few variables suffice for identifying someone. They may be "personal" but one alone is not unique, the combination is

Difficulties

- Difficulties: high dimensional data
  - AOL[3] case

    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'

    $\longrightarrow$ Thelma Arnold identified!

[3] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Difficulties: high dimensional data
  - AOL[3] case
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
  - Netflix (search logs and movie ratings) case
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.

[3] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Difficulties: high dimensional data
  - AOL[3] case
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
  - Netflix (search logs and movie ratings) case
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...

[3] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Difficulties: high dimensional data
  - AOL[3] case
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
  - Netflix (search logs and movie ratings) case
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...

- A large number of variables are needed for identifying someone. The combination of them is identifying

[3] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Data breaches.
  - See e.g. https://en.wikipedia.org/wiki/Data_breach

Difficulties

- Summary of difficulties:
  highly identifiable data and high dimensional data
  - Ex1: Sickness influenced by studies and commuting distance ?
    Problem: original data + reidentification + inference
    (few highly identifiable variables)
    (similar with high dimensional variable)

Difficulties

- Summary of difficulties:
  highly identifiable data and high dimensional data
  - Ex1: Sickness influenced by studies and commuting distance ?
    Problem: original data + reidentification + inference
    (few highly identifiable variables)
    (similar with high dimensional variable)
  - Ex2: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
    Problem: inference from outcome
    (outcome can allow inference on a sensitive variable)

Difficulties

- Summary of difficulties:
  highly identifiable data and high dimensional data
  - Ex3: Driving behavior in the morning
    - ⋆ Automobile manufacturer uses (data from vehicles)
    - ⋆ Data: First drive after 6:00am
      (GPS origin + destination, time) × 30 days
    - ⋆ No "personal data", is this ok?: NO!!!:
    - ⋆ How many cars from your home to your work?
      Are you exceeding the speed limit? Are you visiting a psychiatric clinic every tuesday?

Difficulties

- Summary of difficulties:
  highly identifiable data and high dimensional data
  - Ex3: Driving behavior in the morning
    - ⋆ Automobile manufacturer uses (data from vehicles)
    - ⋆ Data: First drive after 6:00am
      (GPS origin + destination, time) × 30 days
    - ⋆ No "personal data", is this ok?: NO!!!:
    - ⋆ How many cars from your home to your work?
      Are you exceeding the speed limit? Are you visiting a psychiatric clinic every tuesday?

  Problem: original data + reidentification + inference
  + legal implications of acquired knowledge (?)

Difficulties

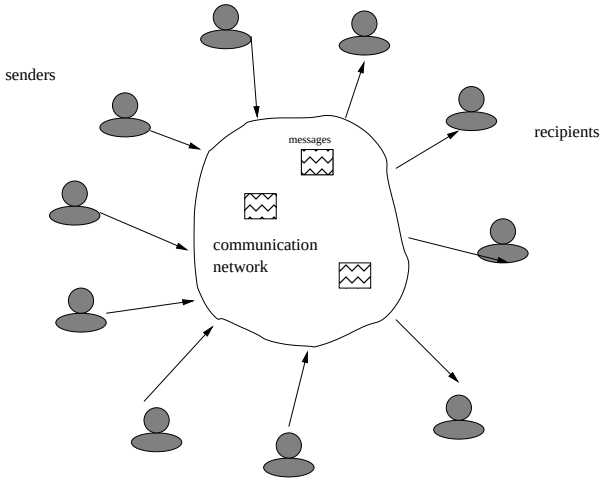
- Data privacy is "impossible", or not? challenging
  - Privacy vs. utility
  - Privacy vs. security
  - Computationally feasible

Terminology

Terminology

- Attacker, adversary, intruder
  - the set of entities working against some protection goal
  - increase their knowledge (e.g., facts, probabilities, . . .) on the items of interest (IoI) (senders, receivers, messages, actions)

In a communication network with senders (actors) and receivers (actees)



Terminology

- Anonymity set. Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set. That is, not distinguishable!

Terminology

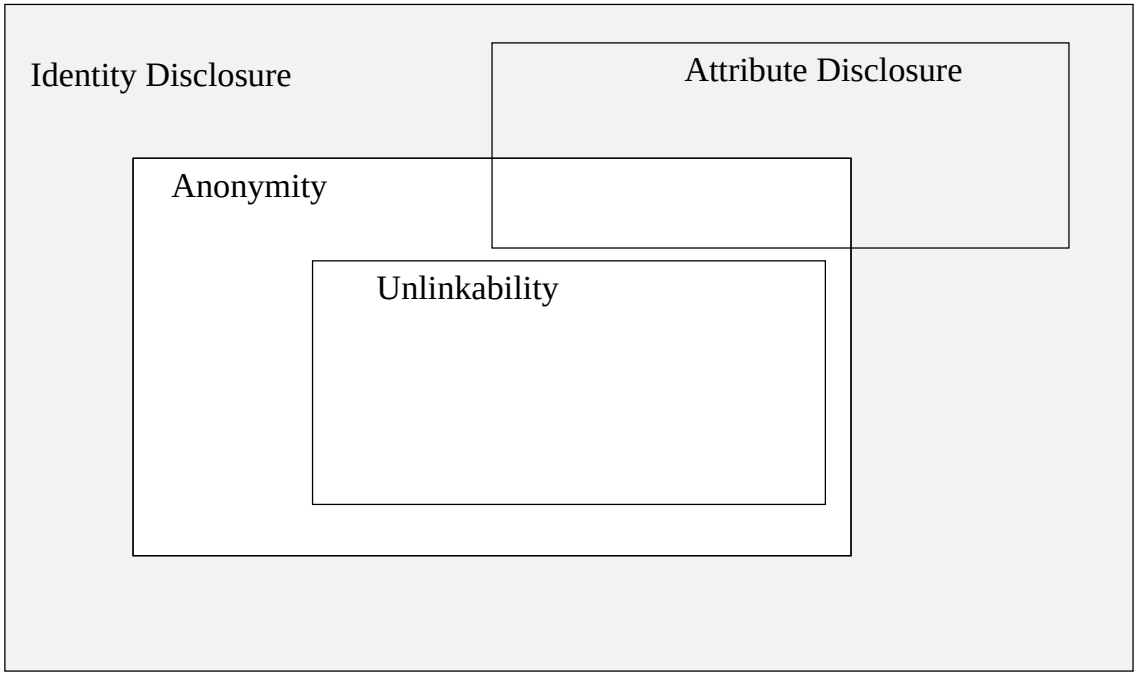
- Anonymity set. Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set. That is, not distinguishable!

- Unlinkability. Unlinkability of two or more IoI, the attacker cannot sufficiently distinguish whether these IoIs are related or not.
  $\Rightarrow$ Unlinkability with the sender implies anonymity of the sender.

  - Linkability but anonymity. E.g., an attacker links all messages of a transaction, due to timing, but all are encrypted and no information can be obtained about the subjects in the transactions: anonymity not compromised.
    (region of the anonymity box outside unlinkability box)

Terminology

- Concepts:
  - Unlinkability implies anonymity



Terminology

- Disclosure. Attackers take advantage of observations to improve their knowledge on some confidential information about an IoI.
  $\Rightarrow$ SDC/PPDM: Observe DB, Δ knowledge of a particular subject (the respondent in a database)

Terminology

- Disclosure. Attackers take advantage of observations to improve their knowledge on some confidential information about an IoI.
  $\Rightarrow$ SDC/PPDM: Observe DB, Δ knowledge of a particular subject (the respondent in a database)
  - Identity disclosure (entity disclosure). Linkability. Finding Mary in the database.
  - Attribute disclosure. Increase knowledge on Mary's salary.
    also: learning that someone is in the database, although not found.

Terminology

- Disclosure. Discussion.
  - Identity disclosure. Avoid.
  - Attribute disclosure. A more complex case. Some attribute disclosure is expected in data mining.

    At the other extreme, any improvement in our knowledge about an individual could be considered an intrusion. The latter is particularly likely to cause a problem for data mining, as the goal is to improve our knowledge. (J. Vaidya et al., 2006, p. 7.

Terminology

- Identity disclosure vs. attribute disclosure
  - identity disclosure implies attribute disclosure (usual case)
    Find record $(HYU, Tarragona, 58)$, learn variable (*Heart Attack*)

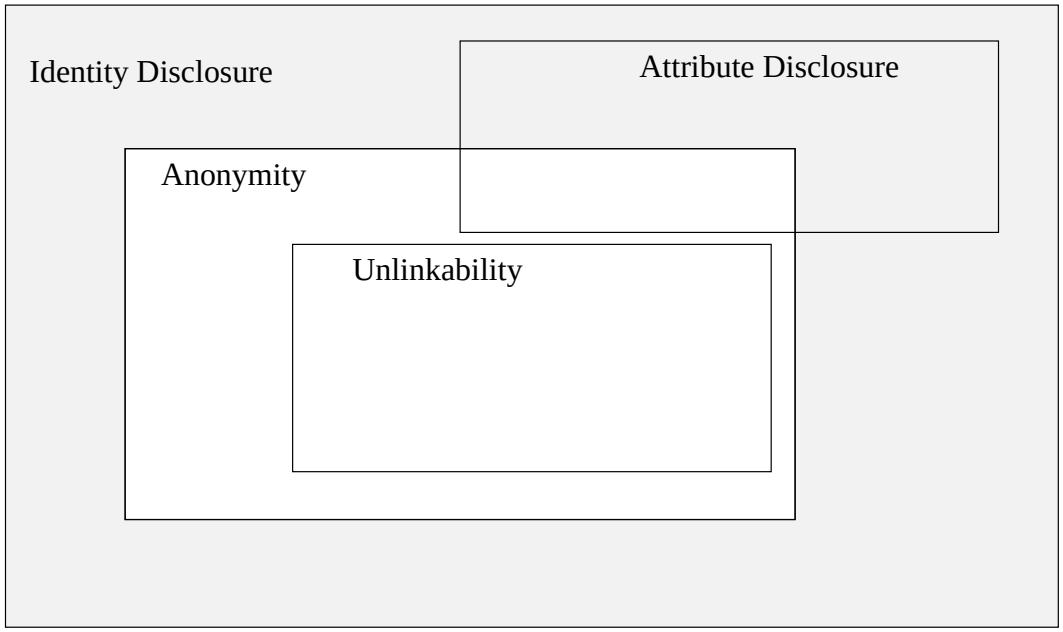
Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	58	Heart attack

  - Identity disclosure without attribute disclosure. Use all attributes
  - Attribute disclosure without identity disclosure. k-anonymity
    $(ABD, Barcelona, 30)$ not reidentified but learn *Cancer*

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS

Terminology

- Identity disclosure and anonymity are exclusive.
  - Identity disclosure implies non-anonymity
  - Anonymity implies no identity disclosure.



Terminology

- Undetectability and unobservability
  - Undetectability of an IoI. The attacker cannot sufficiently distinguish whether IoI exists or not.
    E.g. Intruders cannot distinguish messages from random noise
    $\Rightarrow$ Steganography (embed undetectable messages)

Terminology

- Undetectability and unobservability
  - Undetectability of an IoI. The attacker cannot sufficiently distinguish whether IoI exists or not.
    E.g. Intruders cannot distinguish messages from random noise
    $\Rightarrow$ Steganography (embed undetectable messages)
  - Unobservability of an IoI means
    - $\star$ undetectability of the IoI against all subjects uninvolved in it and
    - $\star$ anonymity of the subject(s) involved in the IoI even against the other subject(s) involved in that IoI.

Unobservability pressumes undetectability but at the same time it also pressumes anonymity in case the items are detected by the subjects involved in the system. From this definition, it is clear that unobservability implies anonymity and undetectability.

Transparency

Transparency

- Transparency
  - DB is published: give details on how data has been produced. Description of any data protection process and parameters
  - Positive effect on data utility. Use information in data analysis.
  - Negative effect on risk. Intruders use the information to attack.

Example. DB masking using additive noise: $X' = X + \epsilon$ with ϵ s.t. $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given constant k then, $Var(X') = Var(X) + kVar(X) = (1 + k)Var(X)$

Transparency

- The transparency principle in data privacy[4]

  Given a privacy model, a masking method should be compliant with this privacy model even if everything about the method is public knowledge. (Torra, 2017, p17)

[4] Similar to the Kerckhoffs's principle (Kerckhoffs, 1883) in cryptography: a cryptosystem should be secure even if everything about the system is public knowledge, except the key

Transparency

- The transparency principle in data privacy[4]

  Given a privacy model, a masking method should be compliant with this privacy model even if everything about the method is public knowledge. (Torra, 2017, p17)

- Transparency a requirement of Trustworthy AI. Related to three elements: traceability, explicability (why decisions are made), and comunication (distinguish AI systems from humans). Transparency in data privacy relates to traceability.

[4] Similar to the Kerckhoffs's principle (Kerckhoffs, 1883) in cryptography: a cryptosystem should be secure even if everything about the system is public knowledge, except the key

Privacy by design

Privacy by design

- Privacy by design (Cavoukian, 2011)
  - Privacy "must ideally become an organization's default mode of operation" (Cavoukian, 2011) and thus, not something to be considered a posteriori. In this way, privacy requirements need to be specified, and then software and systems need to be engineered from the beginning taking these requirements into account.
  - *In the context of developing IT systems, this implies that privacy protection is a system requirement that must be treated like any other functional requirement. In particular, privacy protection (together with all other requirements) will determine the design and implementation of the system* (Hoepman, 2014)

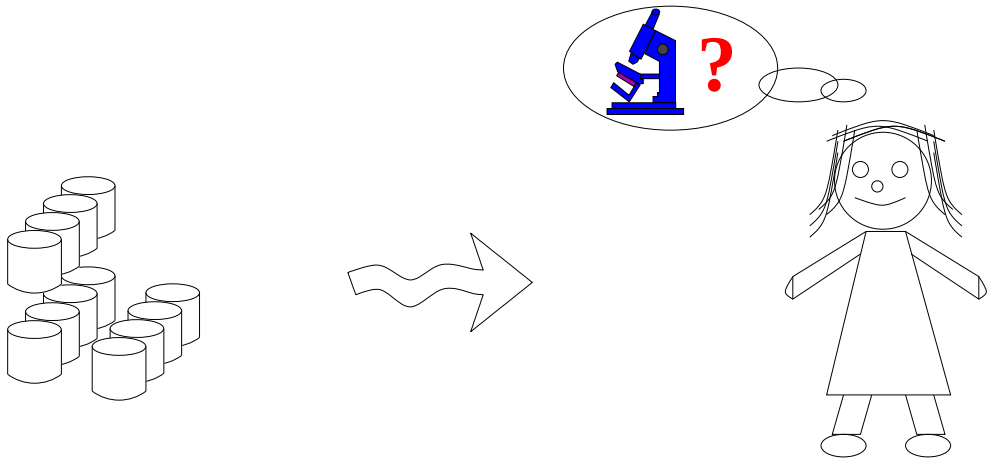
Privacy by design

- Privacy by design principles (Cavoukian, 2011)

1. Proactive not reactive; Preventative not remedial.
2. Privacy as the default setting.
3. Privacy embedded into design.
4. Full functionality – positive-sum, not zero-sum.
5. End-to-end security – full lifecycle protection.
6. Visibility and transparency – keep it open.
7. Respect for user privacy – keep it user-centric.

Privacy models

Privacy models



Privacy models

Privacy models. A computational definition for privacy. Examples.

Privacy models

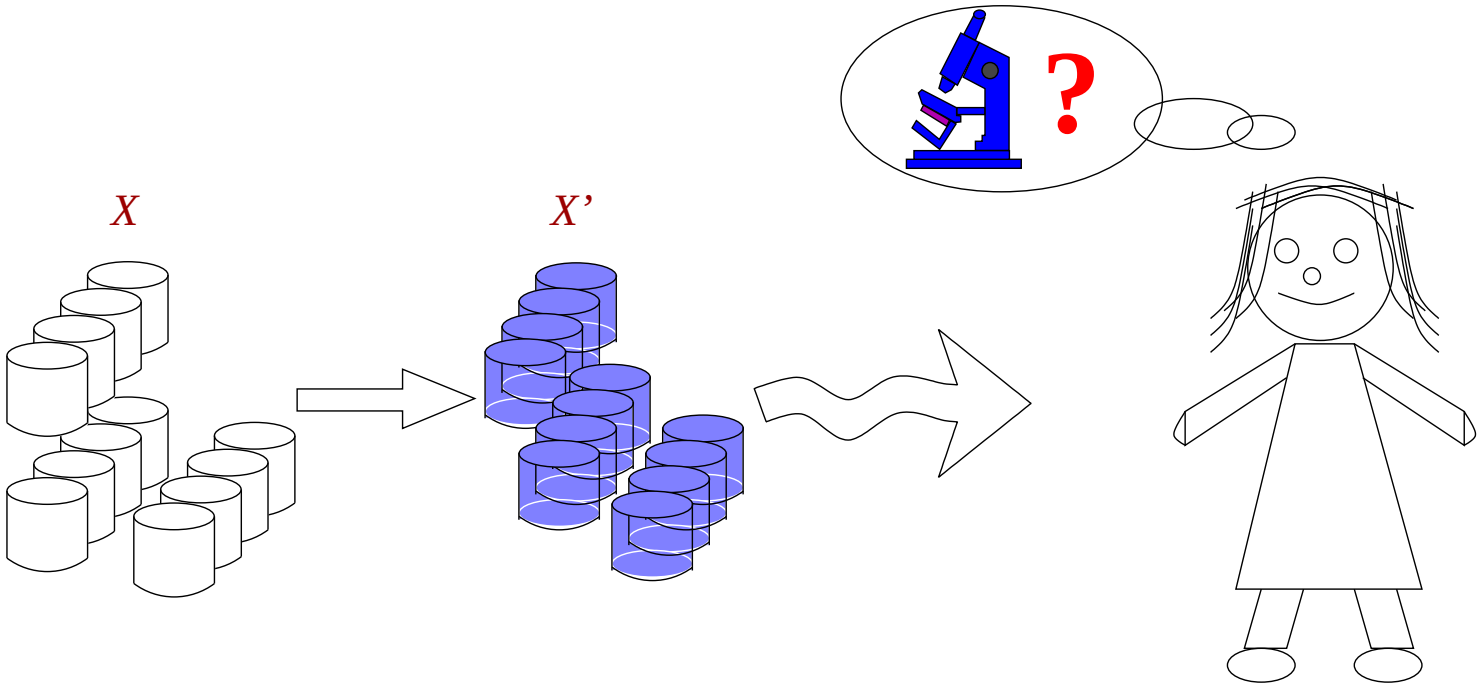
Privacy models. A computational definition for privacy. Examples.

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Result privacy.** We want to avoid some results when an algorithm is applied to a database.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.

Privacy models

Privacy models. A computational definition for privacy. Examples.

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Result privacy.** We want to avoid some results when an algorithm is applied to a database.



Privacy models

- Difficulties: naive anonymization does not work
  - (Sweeney, 1997; 2000[5]) on USA population
    - ⋆ 87.1% (216 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, Month and year of birth.
- Difficulties: highly identifiable data and high dimensional data
  - Data from mobile devices:
    - ⋆ two positions can make you unique (home and working place)
  - AOL and Netflix cases (search logs and movie ratings)
  - Similar with credit card payments, shopping carts, search logs, ... (i.e., high dimensional data)

[5] L. Sweeney, Simple Demographics Often Identify People Uniquely, CMU 2000

Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)

Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",

    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$

    is this ok ?

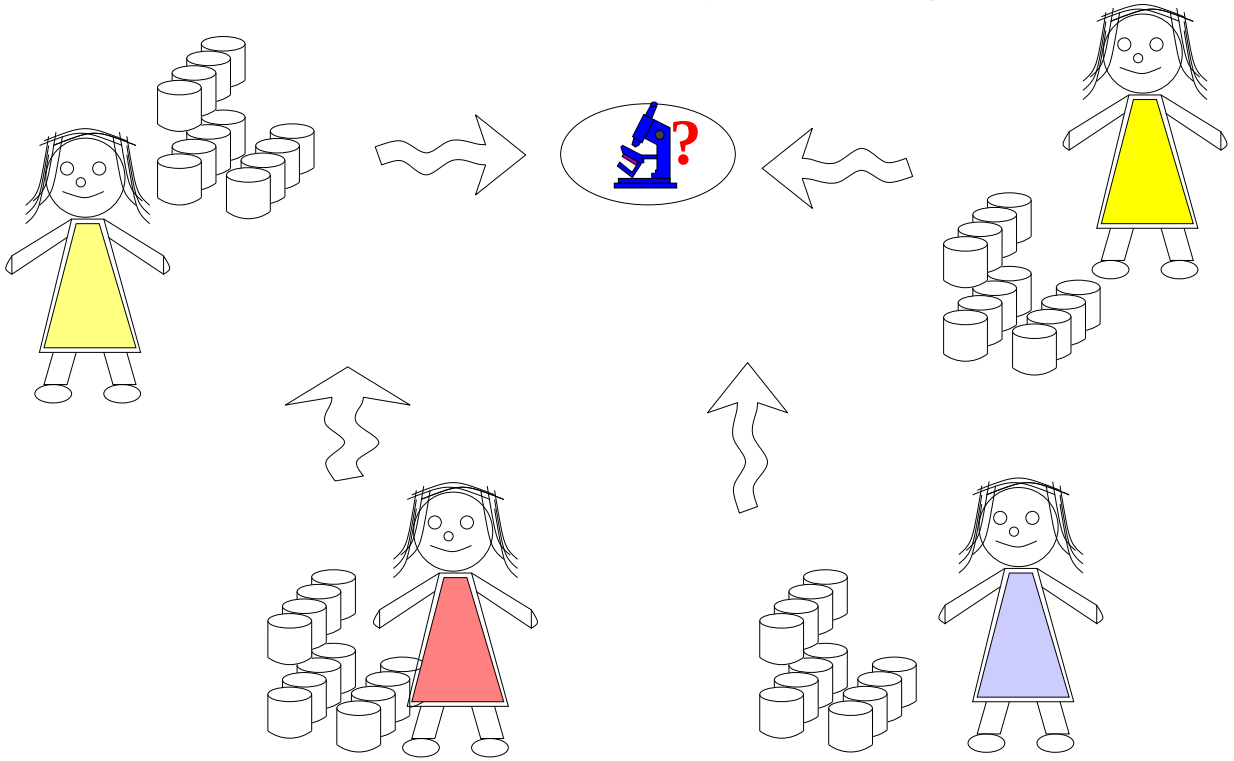
Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$
    is this ok ?
    NO!!:
  - E.g., there is only one student of anthropology living in Enfield.
    (Enfield, Anthropology, Yes)

Privacy models

Privacy models. A computational definition for privacy. Examples.

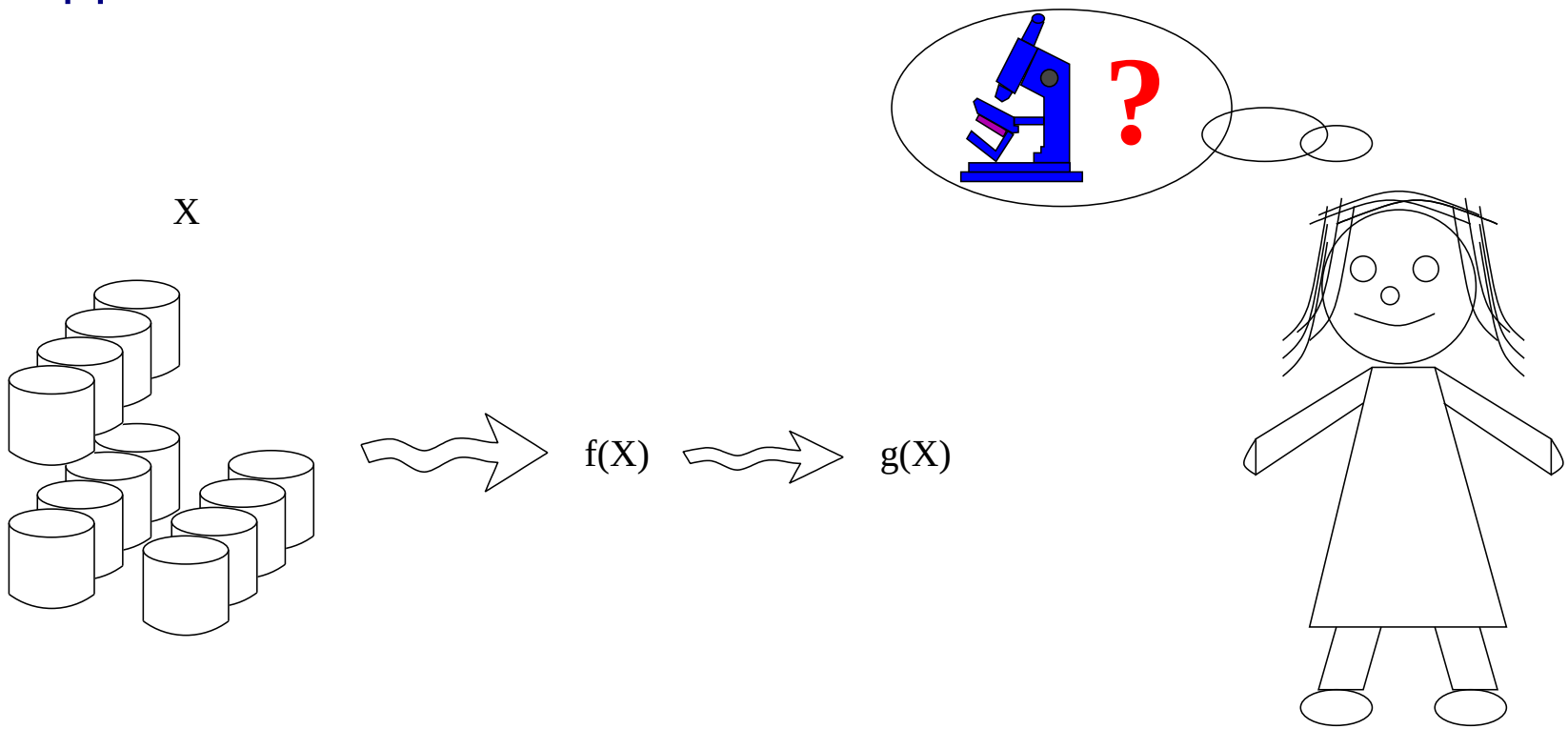
- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.



Privacy models

Privacy models. A computational definition for privacy. Examples.

- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.



Privacy models

- Difficulties. Output of a function can be sensitive. Example 2
  - Mean income of admitted to hospital unit (e.g., psychiatric unit)
  - Mean salary of participants in Alcoholics Anonymous by town

  Is this ok? NO!!

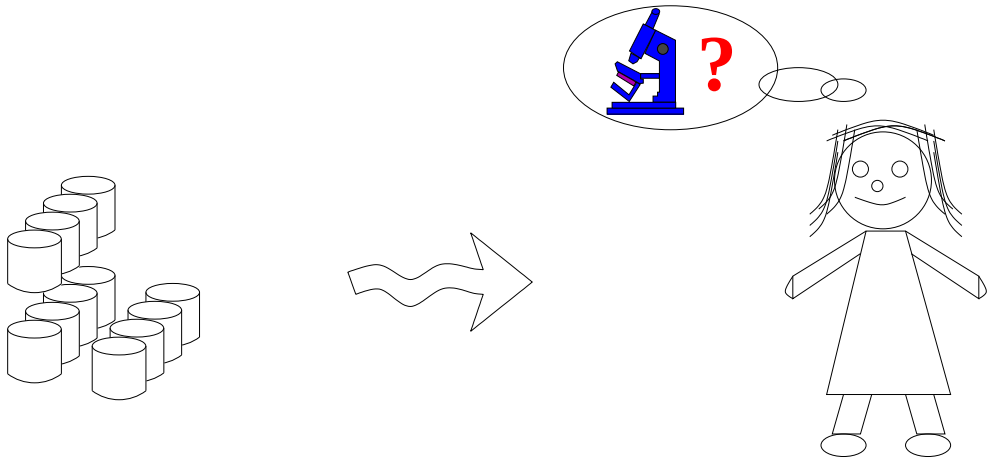
  - disclosure of a rich person in the database

Data privacy mechanisms

Privacy models

Data privacy mechanisms. Classification w.r.t. our knowledge on the computation

- Data-driven or general purpose (*analysis not known*)
  → anonymization / masking methods
- Computation-driven or specific purpose (*analysis known*)
  → cryptographic protocols, differential privacy, integral privacy
- Result-driven (*analysis known: protection of its results*)



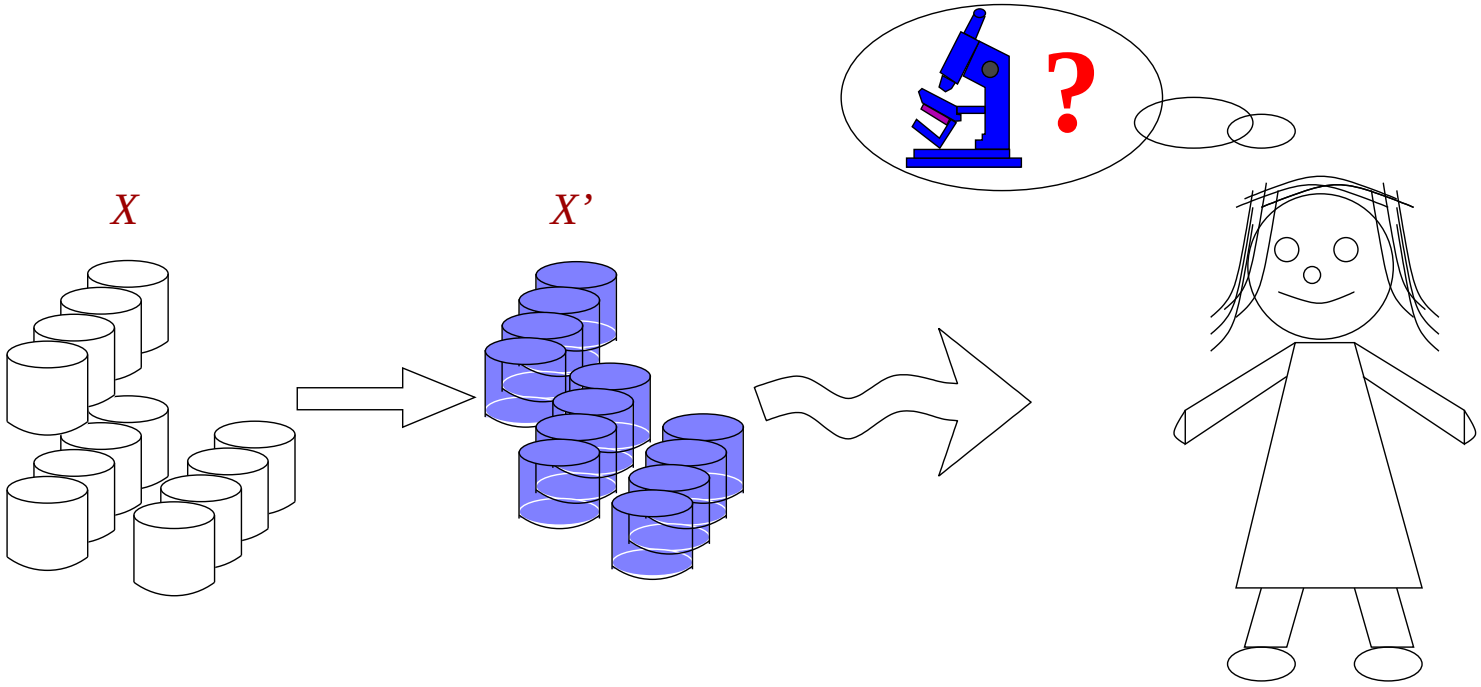
Data privacy mechanisms

Data-driven and general purpose Masking methods

Masking methods

Data-driven or general purpose (*analysis not known*)

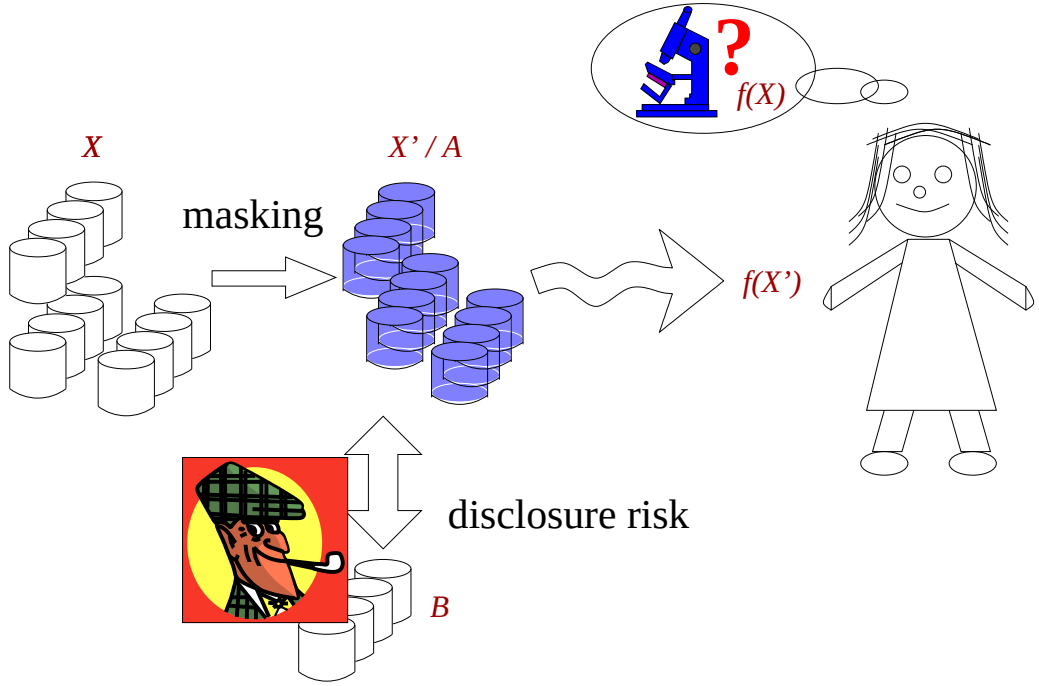
- Privacy model: Reidentification / k-anonymity.
- Privacy mechanisms: Anonymization / masking methods:
  Given a data file X compute a file X' with data of *less quality*.



Masking methods

Data-driven or general purpose (*analysis not known*)

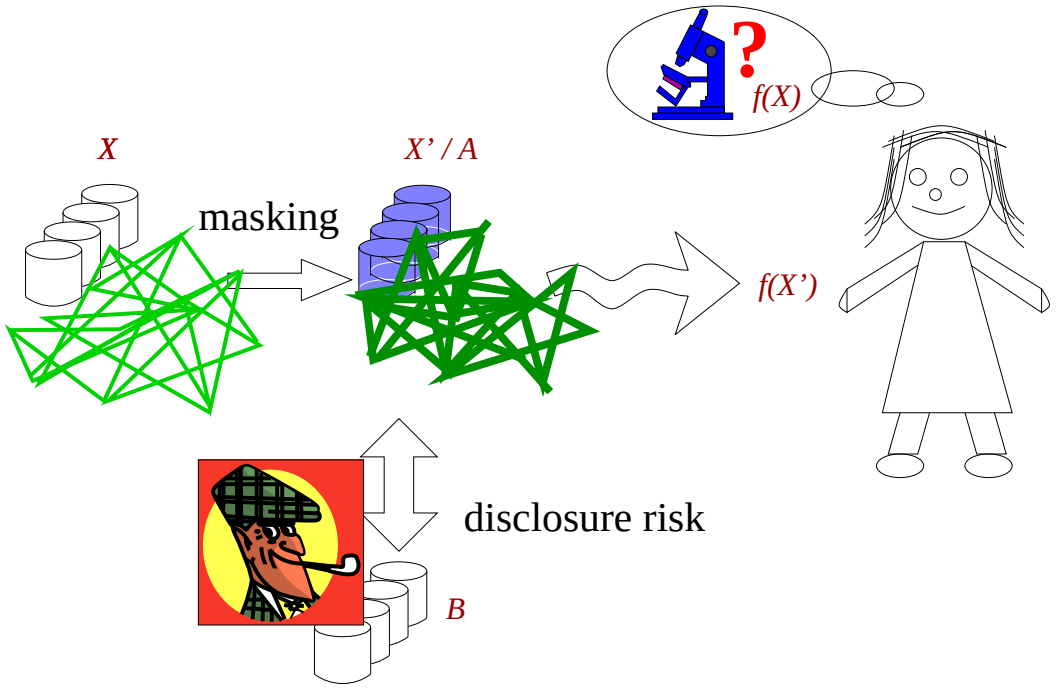
- Privacy model: reidentification / k-anonymity
- Privacy mechanisms: Anonymization / masking methods:
  Given a data file X compute a file X' with data of *less quality*.



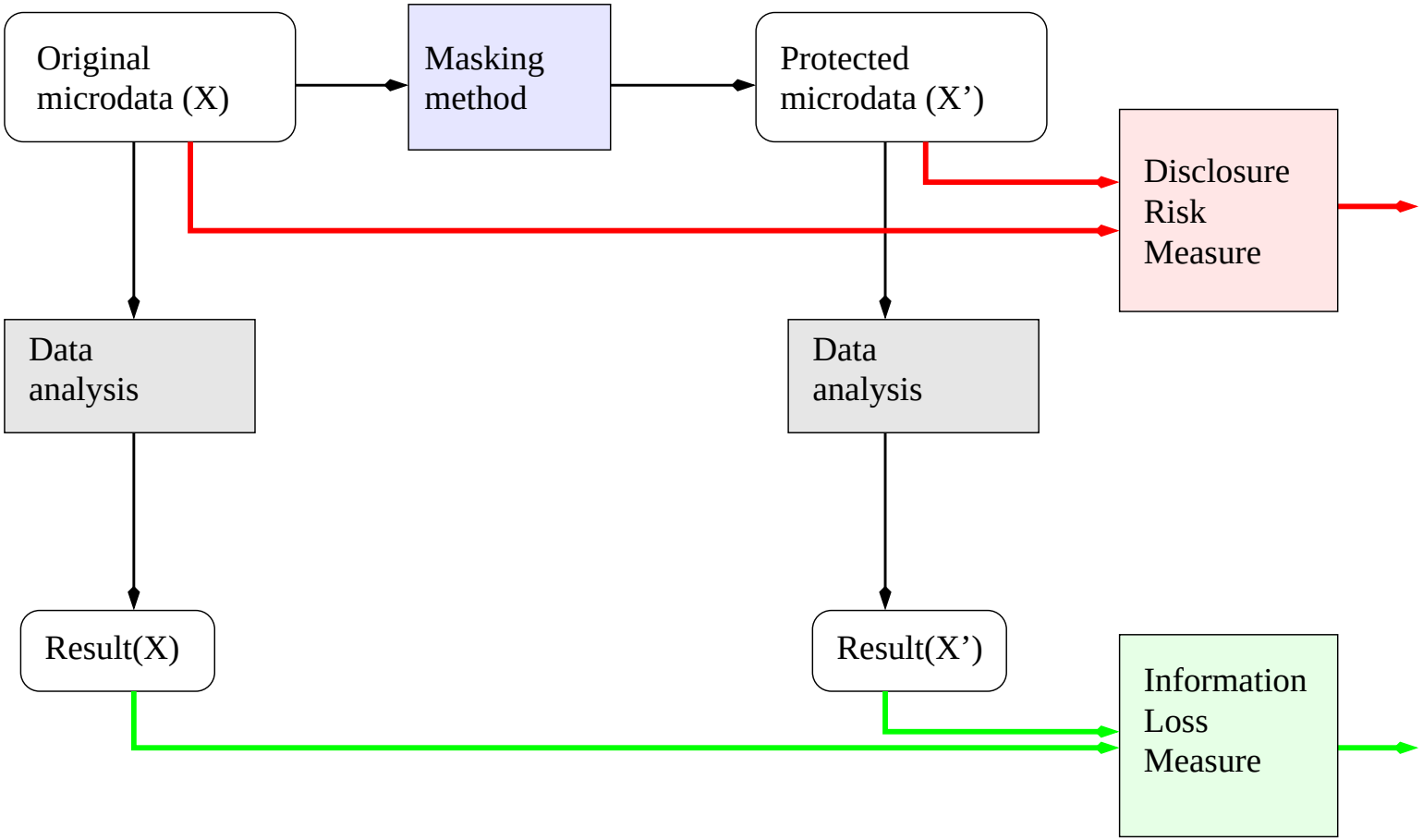
Masking methods

Approach valid for different types of data

- Databases, documents, search logs, social networks, . . .
  (also masking taking into account semantics: wordnet, ODP)



Masking methods



Research questions: (i) masking methods

Masking methods. (anonymization methods) $X' = \rho(X)$

- **Privacy models**
  - **k-anonymity.** Single-objective optimization: utility
  - **Privacy from re-identification.** Multi-objective: trade-off U/Risk
- **Families of methods**
  - Perturbative. (less quality=erroneous data)

    E.g. noise addition/multiplication, microaggregation, rank swapping
  - Non-perturbative. (less quality=less detail)

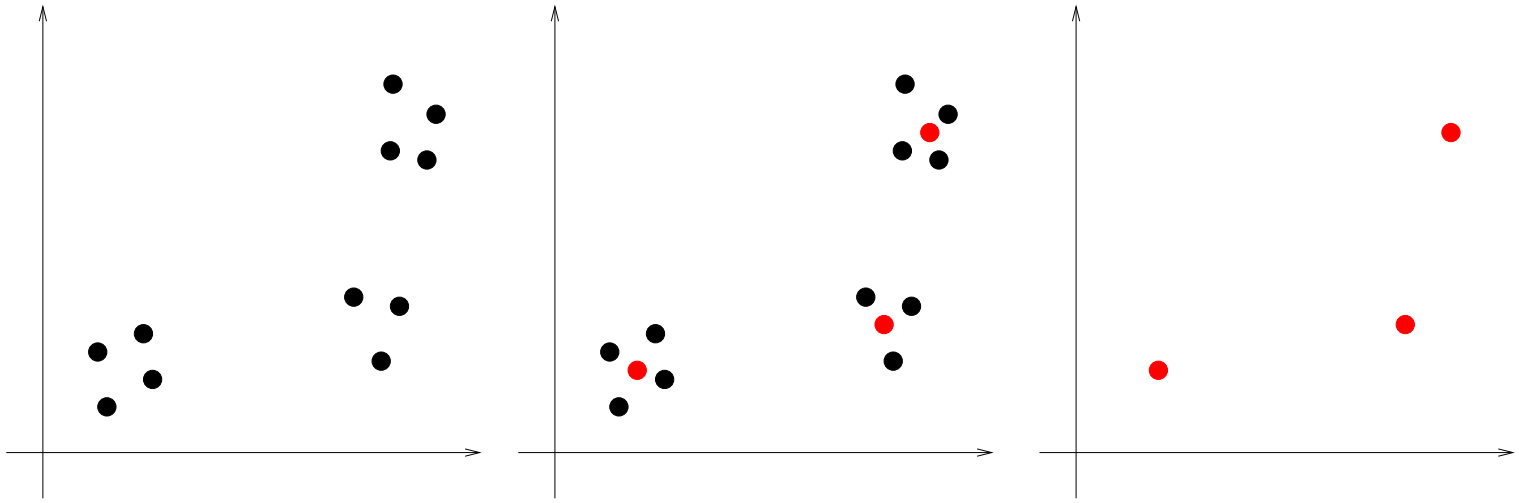
    E.g. generalization, suppression
  - Synthetic data generators. (less quality=not real data)

    E.g. (i) model from the data; (ii) generate data from model

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Microaggregation (k records clusters)

- **Formalization.** ($u_{ij} = 1$ iff x_j in ith cluster; v_i centroid)



Minimize $SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$

Subject to $\sum_{i=1}^{g} u_{ij} = 1$ for all $j = 1, \ldots, n$

$2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \ldots, g$

$u_{ij} \in \{0, 1\}$

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Additive Noise

- **Description.** Add noise into the original file. That is,

$$X' = X + \epsilon,$$

  where ϵ is the noise.
- The simplest approach is to require ϵ to be such that $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given constant k.

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Additive Noise

- **Description.** Add noise into the original file. That is,

$$X' = X + \epsilon,$$

  where ϵ is the noise.
- The simplest approach is to require ϵ to be such that $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given constant k.

Properties:

- It makes no assumptions about the range of possible values for V_i (which may be infinite).
- The noise added is typically continuous and with mean zero, which suits continuous original data well.
- No exact matching is possible with external files.

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. PRAM: Post-Randomization Method

- **Description.**
  - The scores on some categorical variables for certain records in the original file are changed to a different score.
    - $\star$ according to a transition (Markov) matrix
- Properties:
  - PRAM is very general: it encompasses noise addition, data suppression and data recoding.
  - PRAM information loss and disclosure risk largely depend on the choice of the transition matrix.

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Rank swapping

- **Description with parameter p.**
  - Values are ordered in increasing order
    We assume them ordered $x_{ij} \leq x_{lj}$ for all $1 \leq i < l \leq n$
  - Each ranked value x_{ij} is swapped with another ranked value x_{lj} randomly chosen within a restricted range $i < l \leq i + p$
- In applications, each variable is masked independently
- The larger the p, the larger the information loss, and the lower the risk

Research questions: (i) masking methods

Masking methods. $X' = \rho(X)$. Synthetic Data Generators

- **Description.** (partially synthetic data)
  Data: $X|Y$: set of records of a given sample
  Output: $X|Y'$: set of records with Y' a masked version of Y
  1. $M_{X,Y}$:= Build a model of Y in terms of X
  2. $Y' := M_{X,Y}(X)$
  3. Return $(X|Y')$

Need to take attention to disclosure risk. Do not state

> "Since released microdata are synthetic, no real re-identification is possible".

Re-identification can indeed happen if a snooper is able to link an external identified data source with some record in the released dataset using the quasi-identifier attributes: coming up with a correct pair (identifier, confidential attributes) is indeed a re-identification.

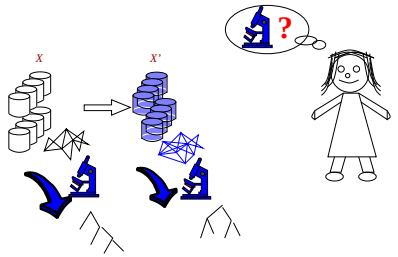
Research questions: (ii) information loss/data utility

Information loss measures. Compare X and X' w.r.t. analysis (f)

$$IL_f(X, X') = divergence(f(X), f(X'))$$

- f: depends on X; generic vs. specific data uses.
  - Statistics, ML: clustering & classification, centrality-graphs, ...
  - For classification using decision trees:

    $accuracy(DT(X))$ vs. $accuracy(DT(X'))$



Research questions: (ii) information loss/data utility

- Typical comparison of methods w.r.t. IL/utility and Risk

			Accuracy, ACC				Area Under Curve, AUC			
	PIL	DR	DT	NB	k-NN	SVM	DT	NB	k-NN	SVM
Original	0.00%	100.00%	54.22%	54.78%	53.93%	54.56%	71.60%	73.30%	71.60%	70.30%
Noise, $\alpha = 3$	7.90%	74.56%	54.39%	51.81%	53.36%	54.49%	73.09%	73.41%	71.48%	70.50%
Noise, $\alpha = 10$	24.65%	38.95%	53.67%	51.88%	51.62%	54.37%	73.24%	73.42%	70.55%	70.49%
Noise, $\alpha = 100$	73.94%	4.10%	51.04%	52.21%	48.17%	53.20%	72.06%	73.98%	66.47%	69.50%
MultNoise, $\alpha = 5$	13.50%	50.81%	54.44%	51.90%	52.36%	54.39%	73.51%	73.42%	71.22%	70.50%
MultNoise, $\alpha = 10$	24.81%	24.75%	54.20%	51.76%	54.20%	54.32%	73.15%	73.42%	72.67%	70.41%
MultNoise, $\alpha = 100$	74.29%	0.00%	50.73%	52.12%	50.90%	53.27%	71.00%	73.90%	68.10%	69.52%
RS p-dist, $p = 2$	22.12%	51.12%	53.19%	51.23%	53.99%	54.37%	70.95%	73.24%	74.15%	70.57%
RS p-dist, $p = 10$	29.00%	23.49%	53.55%	51.85%	54.35%	54.18%	71.84%	73.52%	73.17%	70.40%
RS p-dist, $p = 50$	39.96%	7.80%	40.63%	50.56%	37.32%	53.20%	59.24%	73.17%	57.75%	69.50%
CBFS, $k = 5$	39.05%	13.73%	54.56%	51.64%	54.01%	54.54%	74.10%	73.29%	73.26%	70.62%
CBFS, $k = 25$	58.08%	6.65%	53.31%	51.95%	53.05%	54.01%	73.48%	73.10%	74.22%	70.23%
CBFS, $k = 100$	63.55%	4.32%	51.30%	51.59%	53.53%	54.10%	71.16%	73.24%	74.56%	70.31%
CBFS 2-sen, $k = 25$	58.08%	0.55%	53.31%	52.00%	53.05%	54.13%	73.44%	73.10%	74.22%	70.30%
CBFS 3-sen, $k = 25$	73.00%	0.00%	45.00%	42.00%	43.00%	41.00%	62.00%	61.00%	63.00%	60.00%
CBFS 2-div, $k = 25$	61.55%	0.40%	52.72%	51.57%	52.84%	54.37%	72.13%	73.24%	73.09%	70.36%
CBFS 3-div, $k = 25$	86.00%	0.00%	38.00%	39.00%	38.00%	40.00%	60.00%	61.00%	62.00%	63.00%
IPSO $g = 2$	65.09%	1.66%	52.81%	51.52%	50.11%	53.39%	72.36%	73.61%	68.06%	69.66%
IPSO $g = 3$	58.93%	4.93%	51.45%	51.09%	49.87%	52.41%	69.58%	73.22%	68.24%	68.81%
IPSO $g = 4$	58.56%	1.81%	52.05%	51.23%	50.68%	52.52%	70.41%	73.22%	68.52%	69.00%

Abalone (4177 records, 9 attr, 3 classes) w/ different SDC perturbation methods[6].

[6] Herranz, Matwin, Nin, Torra (2010) Classifying data from protected statistical datasets. C&S.

Research questions: (ii) information loss/data utility

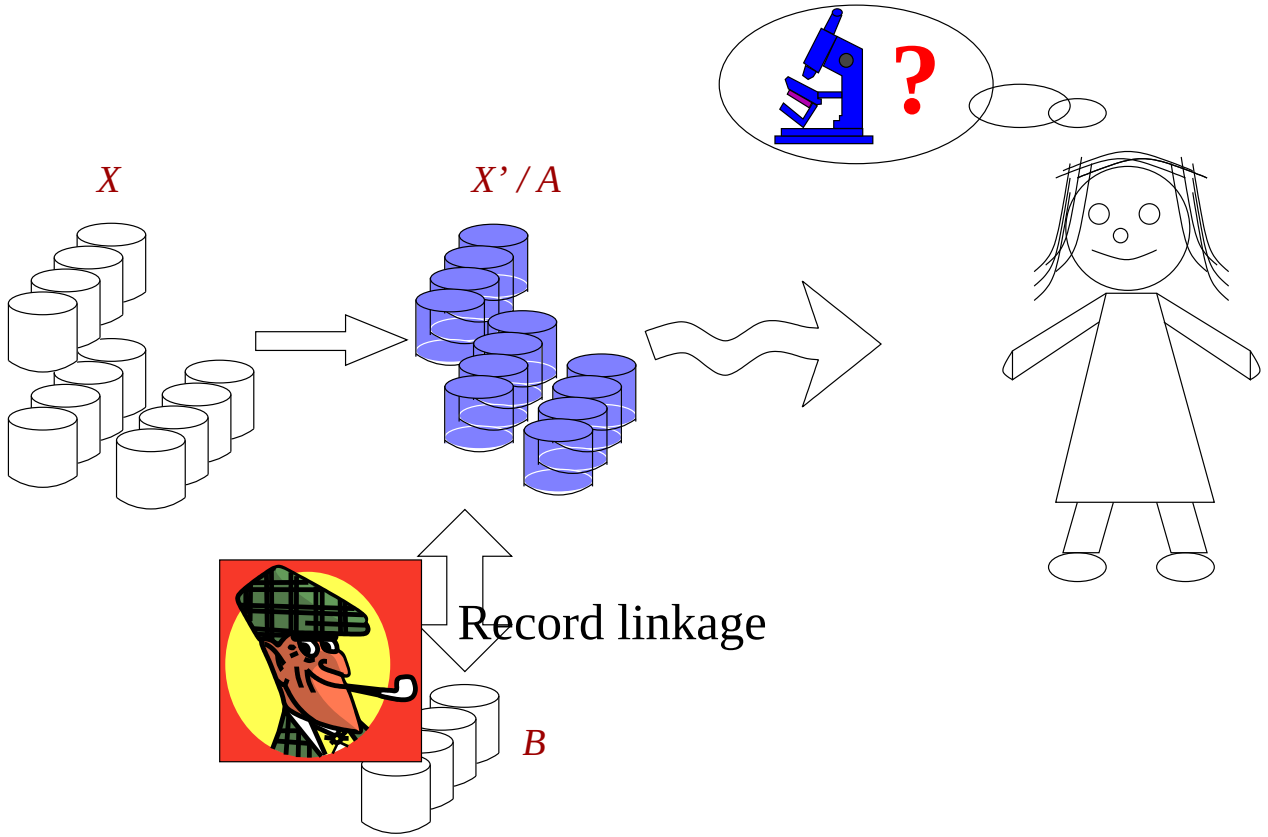
ML models, accuracy and masking methods

- Masking methods: not always equivalent to a loss of accuracy

  There are cases in which the performance is even improved. Aggarwal and Yu (2004) report that 'in many cases, the classification accuracy improves because of the noise reduction effects of the condensation process'. The same was concluded in [Sakuma and Osame, 2017] for recommender systems: 'we observe that the prediction accuracy of recommendations based on anonymized ratings can be better than those based on non- anonymized ratings in some settings'. [Torra, 2017]

Research questions: (iii) disclosure risk assessment

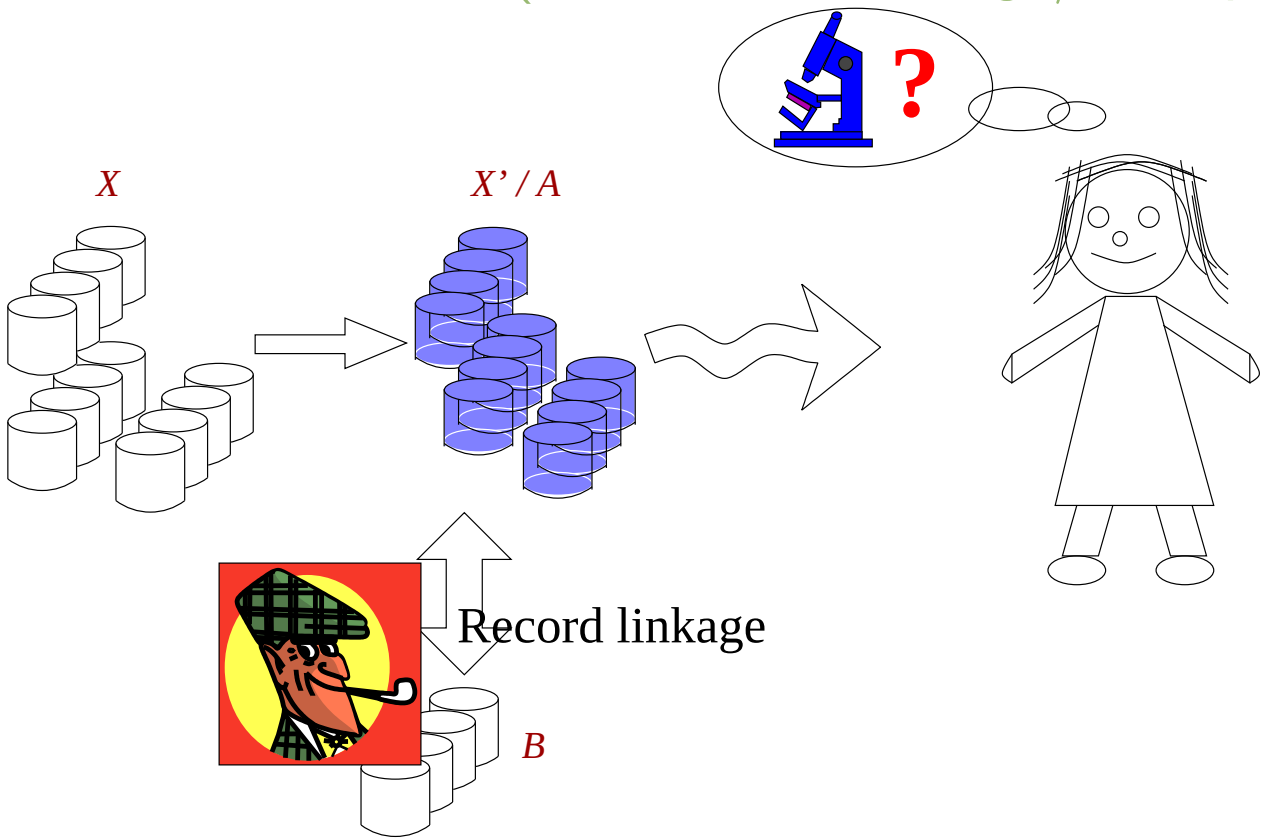
- **Privacy from re-identification**. Identity disclosure[7]. Scenario:
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)



[7] Identity disclosure vs. attribute disclosure: Finding Alice in DB vs. Δ knowledge on Alice's salary

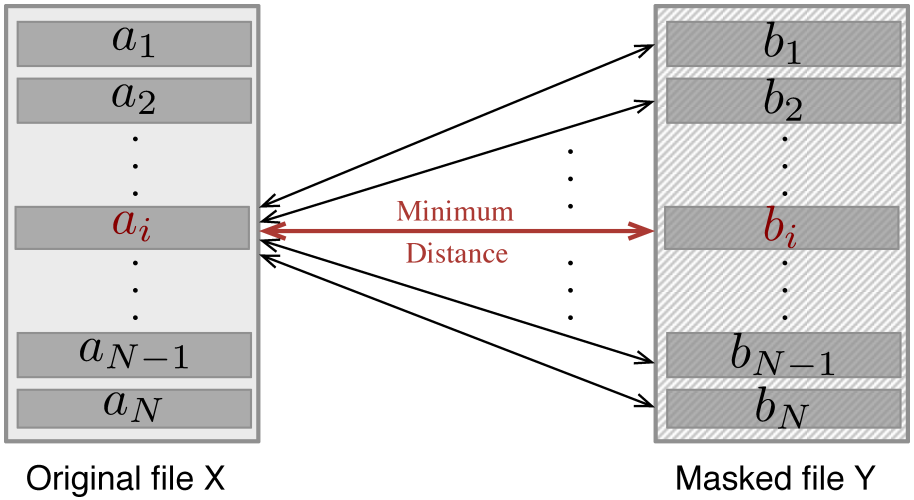
Research questions: (iii) disclosure risk assessment

- **Privacy from re-identification**. Worst-case scenario (maximum knowledge) to give upper bounds of risk:
  - transparency attacks (information on how data has been protected)
  - largest data set (original data)
  - best re-identification method (best record linkage/best parameters)



Research questions: (iii) disclosure risk assessment

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
    - ⋆ Goal: as many correct reidentifications as possible:
      for each record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j

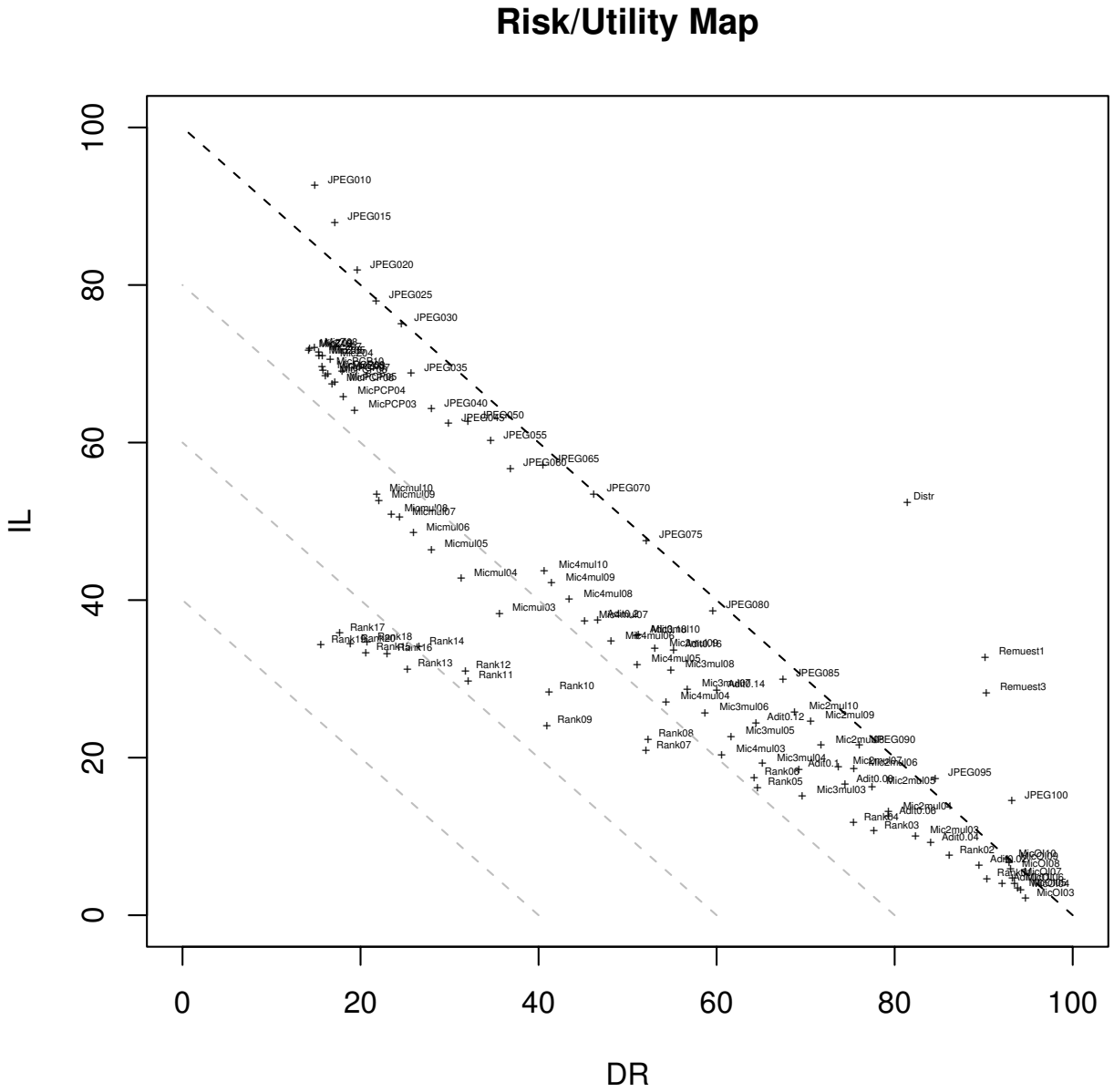


- $d(a_i, b_j)$ as average/sum of attribute/variable distances

$$\mathbb{C}_p(diff_1(a_i, b_j), \ldots, diff_n(a_i, b_j))$$

Research questions: (i)+(ii)+(iii) visualization

- **Comparing masking methods**. Information loss and risk



Data privacy mechanisms
Computation-driven and specific purpose

Computation-driven: "Whose privacy" perspective

Respondent and owner privacy

- Data-driven or general-purpose
- Computation-driven or specific-purpose (Ch. 3.4)
  - Single database: differential privacy (Ch. 3.4.1)
  - Multiple databases:
    - Centralized approach: trusted third party (Ch. 3.4.2)
    - Distributed approach: secure multiparty computation (Ch. 3.4.2)
- Result-driven

Data privacy mechanisms

Computation-driven Differential privacy

Differential privacy

- Computation-driven/single database
  - Privacy model: differential privacy[8]
  - We know the function/query to apply to the database: f
- Example:

  compute the mean of the attribute salary of the database for all those living in Town.

[8] There are other models as e.g. query auditing (determining if answering a query can lead to a privacy breach), and integral privacy

Differential privacy

- Differential privacy (Dwork, 2006).
  - Motivation:
    - ⋆ the result of a query should not depend on the presence (or absence) of a particular individual
    - ⋆ the impact of any individual in the output of the query is limited

  differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis (Dwork, 2006)

Differential privacy

- Mathematical definition of differential privacy
  (in terms of a probability distribution on the range of the function/query)
  - A function K_q for a query q gives ϵ-differential privacy if for all data sets D_1 and D_2 differing in at most one element, and all $S \subseteq Range(K_q)$,

$$\frac{Pr[K_q(D_1) \in S]}{Pr[K_q(D_2) \in S]} \leq e^{\epsilon}.$$

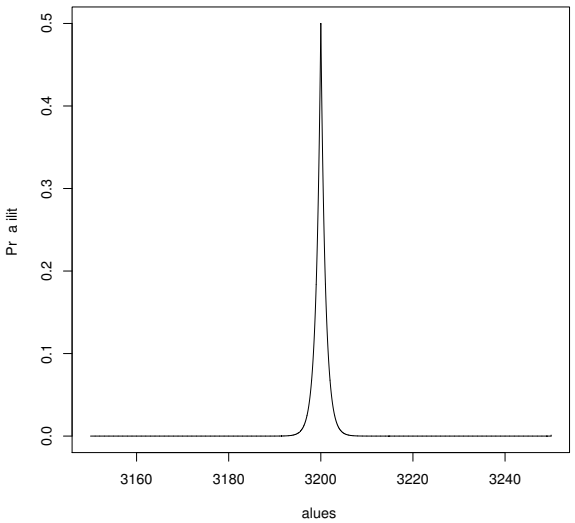
    (with 0/0=1) or, equivalently,

$$Pr[K_q(D_1) \in S] \leq e^{\epsilon} Pr[K_q(D_2) \in S].$$

- ϵ is the level of privacy required (*privacy budget*).
  The smaller the ϵ, the greater the privacy we have.

Differential privacy

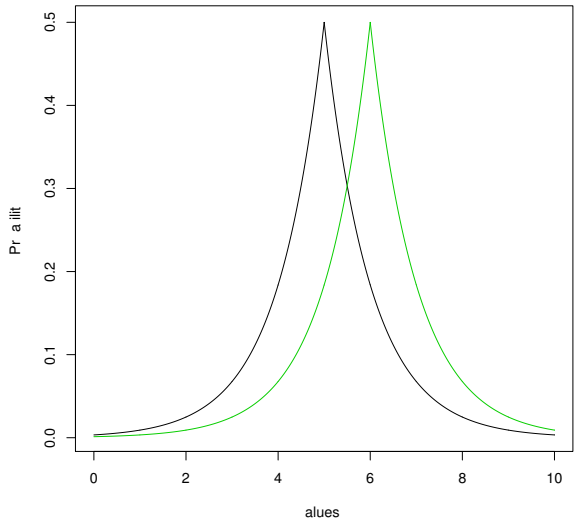
- Differential privacy[9]
  - A function K_q for a query q gives ϵ-differential privacy if . . .
    - $K_q(D)$ ~~is a constant. E.g.,~~
      $$~~K_q(D) = 0~~$$
    - $K_q(D)$ is a randomized version of $q(D)$:
      $$K_q(D) = q(D) + \textit{and some appropriate noise}$$



[9] Self-proclaimed the *de facto standard* for data privacy

Differential privacy

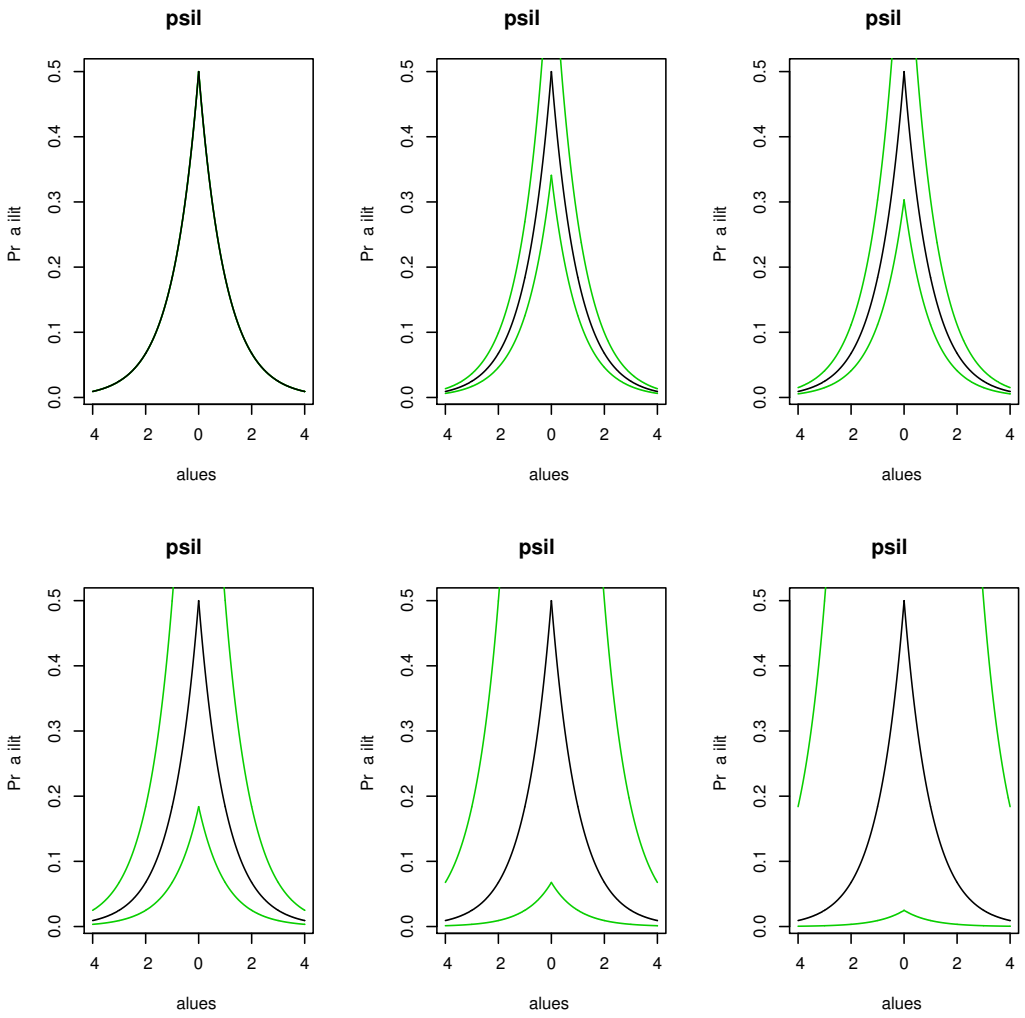
- Differential privacy
  - $K_q(D)$ for a query q is a randomized version of $q(D)$
    - ⋆ Given two neighbouring databases D and D'
      $K_q(D)$ and $K_q(D')$ should be similar enough . . .
  - Example with $q(D) = 5$ and $q(D') = 6$ and adding a Laplacian noise $L(0, 1)$



  - Let us compare different ϵ for noise following $L(0, 1)$ . . .

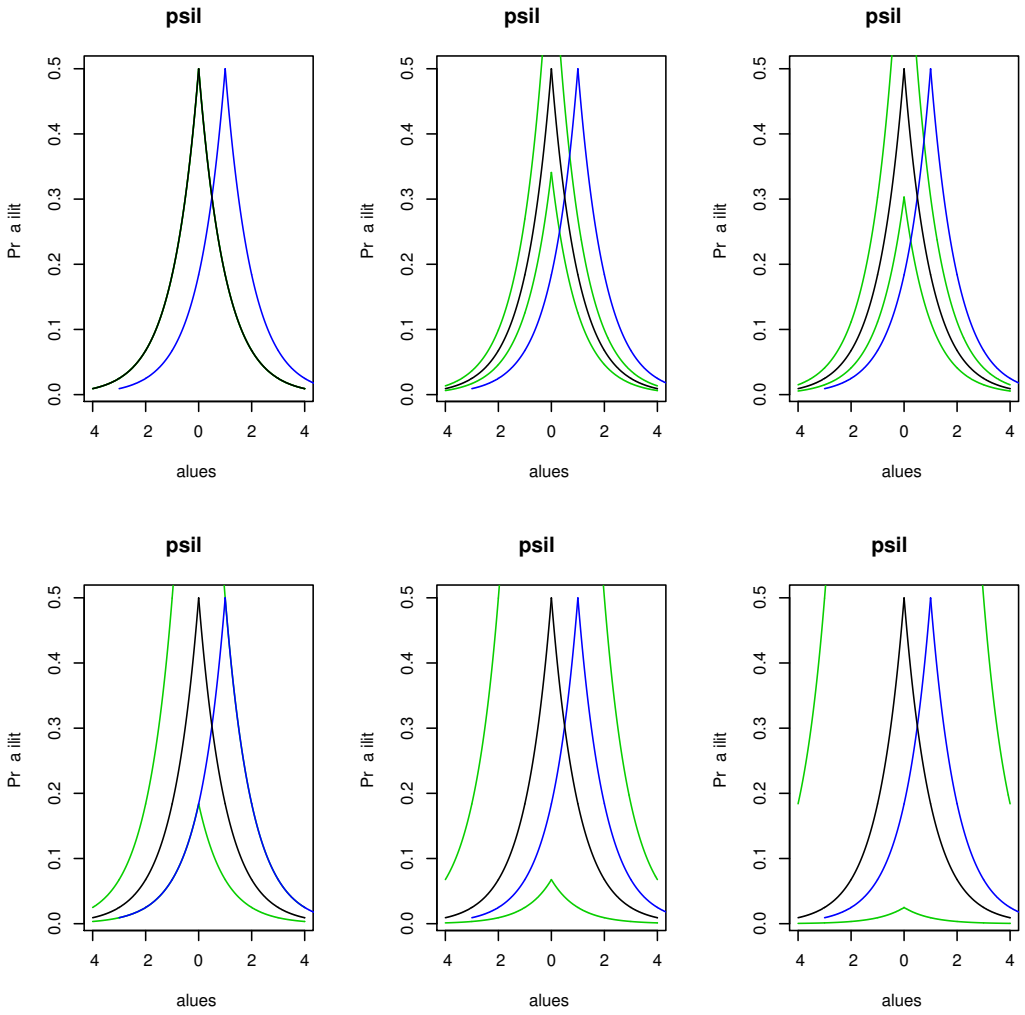
Differential privacy: comparing ϵ for $L(0,1)$

Original $L(0,1)$ and $L(0,1)/e^{\epsilon}, L(0,1) \cdot e^{\epsilon}$



Differential privacy: Accepting 0+2? (using ϵ,$L(0,1)$)

Can $0 + 2$ be acceptable ? I.e., with a distribution similar enough?



Differential privacy

- These examples use the Laplace distribution $L(\mu, b)$.
  - I.e., probability density function:

$$f(x|\mu, b) = \frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right)$$

where

⋆ μ: location parameter

⋆ b: scale parameter (with $b > 0$)

- Properties
  - When $b = 1$, the function for $x > 0$ corresponds to the exponential distribution scaled by $1/2$.
  - Laplace has fatter tails than the normal distribution
  - When $\mu = 0$, for all translations $z \in \mathbb{R}$, $h(x+z)/h(x) \leq exp(|z|)$.

Differential privacy

- Implementation of differential privacy for a numerical query.
  - $K_q(D)$ is a randomized version of $q(D)$:
    $$K_q(D) = q(D) + \textit{and some appropriate noise}$$
  - What is $\textit{and some appropriate noise}$?
- Sensitivity of a query
  - Let $\mathcal{D}$ denote the space of all databases; let $q : \mathcal{D} \to \mathbb{R}^d$ be a query; then, the sensitivity of q is defined

$$\Delta_{\mathcal{D}}(q) = \max_{D,D' \in \mathcal{D}} ||q(D) - q(D')||_1.$$

where $||\cdot||_1$ is the L_1 norm, that is, $||(a_1, \ldots, a_d)||_1 = \sum_{i=1}^{d} |a_i|$.

- Definition essentially meaningful when data has upper & lower bounds

Differential privacy

- Implementation of differential privacy: The case of the mean.
  - Sensitivity of the mean:

$$\Delta_{\mathcal{D}}(mean) = (max - min)/S$$

    where $[min, max]$ is the range of the attribute, and S is the minimal cardinality of the set.
    - ⋆ If no assumption is made on the size of S: $\Delta_{\mathcal{D}}(mean) = (max - min)$
  - Parameter ϵ:

    (Lee, Clifton, 2011) recommend $\epsilon = 0.3829$ for the mean

Differential privacy

- Implementation of differential privacy for a numerical query.
  - Differential privacy via noise addition to the true response
  - Noise following a Laplace distribution $L(0, b)$ with mean equal to zero and scale parameter $b = \Delta(q)/\epsilon$.
    ($\Delta(q)$ is the sensitivity of the query)
  - **Algorithm** Differential privacy:
    - ⋆ **Input:** D: Database; q: query; ϵ: parameter of differential privacy;
    - ⋆ **Output:** Answer to the query q satisfying ϵ-differential privacy
    - ⋆ $a := q(D)$ with the original data
    - ⋆ $\Delta_{\mathcal{D}}(q)$:= the sensitivity of the query for a space of databases D
    - ⋆ Generate a random noise r from a $L(0, b)$ where $b = \Delta(q)/\epsilon$
    - ⋆ Return $a + r$

Differential privacy

- Implementation of differential privacy: The case of the mean.
  - Example[10]:
    - ⋆ $D = \{1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000\}$
      $\Rightarrow$ mean $= 3300$
    - ⋆ Adding Ms. Rich's salary 100,000 Eur/month: mean $= 12090{,}90$!
      (a extremely high salary changes the mean significantly)
      $\Rightarrow$ We infer Ms. Rich from Town was attending the unit
  - $\Rightarrow$ Differential privacy to solve this problem

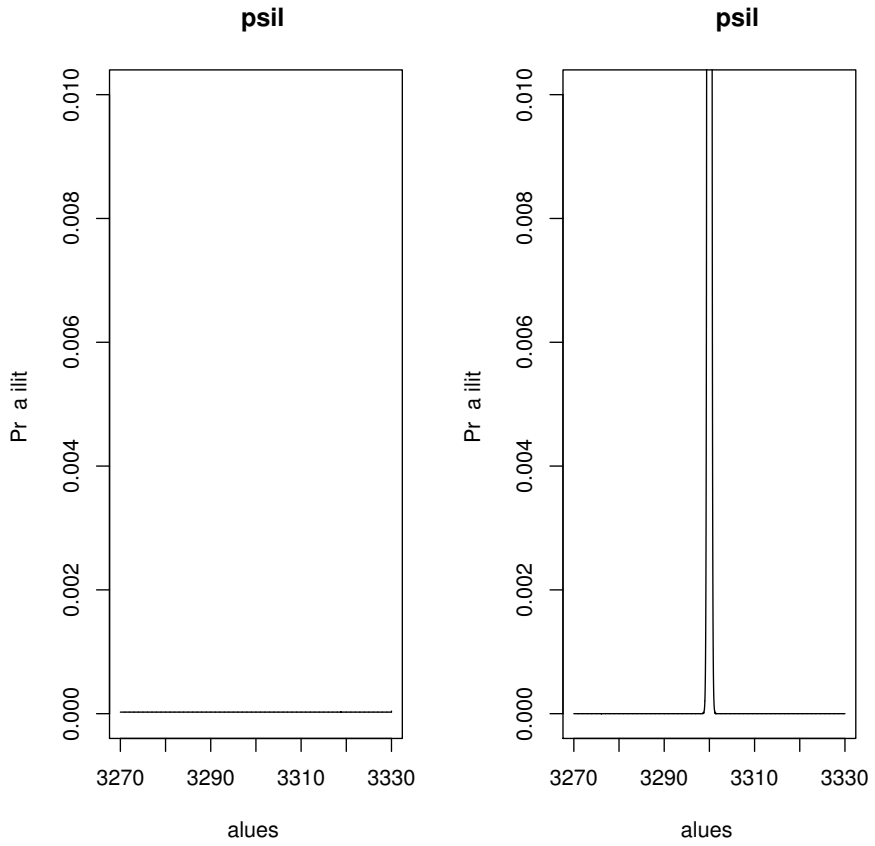
[10] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

Differential privacy

- Implementation of differential privacy: The case of the mean
  - Consider the mean salary
  - Range of salaries $[1000, 100000]$
- Compute for $\epsilon = 1$, assume that at least $S = 5$ records
  - sensitivity $\Delta_{\mathcal{D}}(q) = (max - min)/S = 19800$
  - scale parameter $b = 19800/1 = 19800$
  - For the database: (mean = 3300)
    D={1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000}
  - Output: $K_{mean}(D) = 3300 + L(0, 19800)$
- Compute for $\epsilon = 1$, assume that at least $S = 10^6$ records
  - sensitivity $\Delta_{\mathcal{D}}(q) = (max - min)/S = 0.099$
  - scale parameter $b = 0.099/1 = 0.099$
  - For the database: (mean = 3300)
    D={1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000}
  - Output: $K_{mean}(D) = 3300 + L(0, 0.099)$

Differential privacy: The two distributions

- Comparing
  - (i) $(S = 5, \epsilon = 1)$ $K_{mean}(D) = 3300 + L(0, 19800)$ and
  - (ii) $(S = 10^6, \epsilon = 1)$ $K_{mean}(D) = 3300 + L(0, 0.099)$



Differential privacy

- Laplace mechanism for differential privacy (numerical query)

$$K_q(D) = q(D) + L(0, \Delta(q)/\epsilon)$$

  - **Proposition.** For any function q, the Laplace mechanism satisfies ϵ-differential privacy.

Differential privacy

- Implementation of differential privacy: The case of the mean.
  - "Clamping down" on the output: (McSherry, 2009; Li, Lyu, Su, Yang, 2016 Sections 2.5.3 and 2.5.4)
    - ⋆ The output of a query is within a range $[mn, mx]$ even if data is not. E.g., compute $q(D) = q'_{mn,mx}(mean(D))$ with q' as follows

$$q'_{mn,mx}(x) = \begin{cases} mn & \text{if } x < mn \\ x & \text{if } mn \leq x \leq mx \\ mx & \text{if } mx < x \end{cases}$$

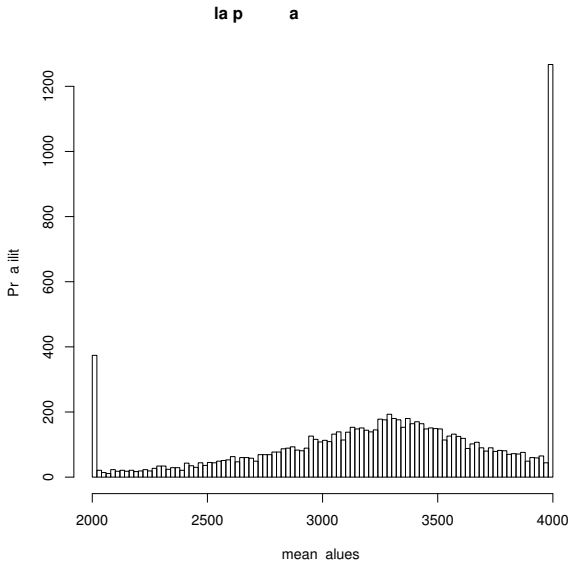
  $\Rightarrow$ we can define ϵ-differential privacy for this query $q(D)$

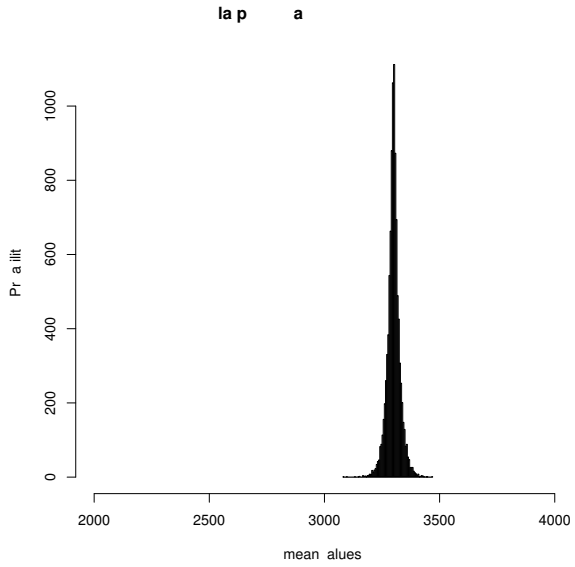
Differential privacy

- Implementation of clamping-down mean
  - Differential privacy via noise addition to the true response
  - Arbitrary size S of the database D (i.e, $S = |D|$)
  - Output in the interval $[mn, mx]$
  - Solution and proof in (Li, Lyu, Su, Yang, 2016 Section 2.5.4)
  - **Algorithm** Differentially private clamping-down mean
    - ⋆ **Input:** D: (one-dimensional) Database; S : $size$; ϵ: parameter of differential privacy; mn,mx: real
    - ⋆ **Output:** A ϵ-differentially private mean
    - ⋆ if $S = 0$ then
      $r :=$ uniform random in $[0, 1]$
      if $r < 1/2exp(-\epsilon/2)$ return mn
      else if $r < 2/2exp(-\epsilon/2)$ return mx
      else return $mn + (mx - mn)(r - exp(-\epsilon/2))/(1 - exp(-\epsilon/2))$
    - ⋆ else return $q'\left(\frac{sum(D)+L(0,(mx-mn)/\epsilon)}{S}\right)$
    - ⋆ end if

Differential privacy

- Implementation of clamping-down mean. Applying it to
  - the interval: [2000, 4000]
  - so, sensitivity $\Delta_{\mathcal{D}}(q) = (max - min) = 2000$
  - and the database: (mean = 3300)
    D={1000, 2000, 3000, 2000, 1000, 6000, 2000, 10000, 2000, 4000}
  - Applying the procedure 10000 times, and ploting the histogram





Differential privacy

- Properties of differential privacy
  - On the ϵ:
    - ⋆ Small ϵ, more privacy, more noise into the solution
    - ⋆ Large ϵ, less privacy, less noise into the solution
  - On the *sensitivity*:
    - ⋆ Small sensitivity, less noise for achieving the same privacy
    - ⋆ Large sensitivity, more noise for achieving the same privacy
  - Discussion here is for a single query (with privacy budget ϵ). Multiple queries (even multiple applications of the same query) need special treatment. E.g., additional privacy budget.
  - Randomness via e.g. Laplace means that any number can be selected. Including e.g. negative ones for salaries. Special treatment may be necessary.
  - Implementations for other type of functions
    - ⋆ The exponential mechanism for non-numerical queries
    - ⋆ Differential privacy for machine learning and statistical models

Data privacy mechanisms
Computation-driven
Centralized approach: trusted third party

Trusted third party

Computation-driven approaches/multiple databases: centralized

- **Example.** Parties $P_1, \ldots, P_n$ own databases $DB_1, \ldots, DB_n$. The parties want to compute a function, say f, of these databases (i.e., $f(DB_1, \ldots, DB_n)$) without revealing unnecessary information. In other words, after computing $f(DB_1, \ldots, DB_n)$ and delivering this result to all P_i, what P_i knows is nothing more than what can be deduced from his DB_i and the function f.
- So, the computation of f has not given P_i any extra knowledge.

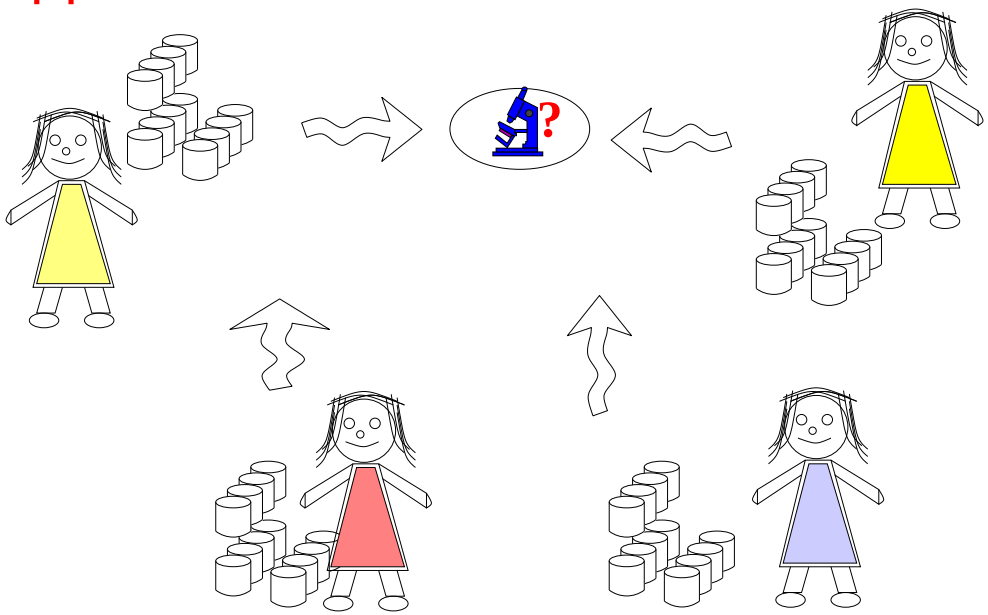
Data privacy mechanisms

Computation-driven
Distributed approach: secure multiparty computation

Secure multiparty computation

Computation-driven approaches/multiple databases: distributed

- The centralized approach as a reference



Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Compute the sum of salaries of 4 people: Aine, Brianna, Cathleen, and Deirdre.
  We denote these salaries by s_1, s_2, s_3, and s_4, respectively.
- Each person's salary is confidential and they do not want to share.
- Define a protocol to compute involving only the 4 people (no trusted third party).
- Assume that the sum lies in the range $[0, n]$.

□ Example with 4 people. Similar method applies with other number of people.

□ We use public-key cryptography. I.e., each party requires two separate keys: a private and a public one. This is also known as asymmetric cryptography.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.
- Brianna decrypts Aine's message with Brianna's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 \ mod \ n$) to Cathleen encrypted with Cathleen's public key.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.
- Brianna decrypts Aine's message with Brianna's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 \ mod \ n$) to Cathleen encrypted with Cathleen's public key.
- Cathleen decrypts Brianna's message with Cathleen's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 \ mod \ n$) to Deirdre encrypted with Deirdre's public key.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.
- Brianna decrypts Aine's message with Brianna's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 \ mod \ n$) to Cathleen encrypted with Cathleen's public key.
- Cathleen decrypts Brianna's message with Cathleen's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 \ mod \ n$) to Deirdre encrypted with Deirdre's public key.
- Deirdre decrypts Cathleen's message with Deirdre's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 + s_4 \ mod \ n$) to Aine encrypted with Aine's public key.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.
- Brianna decrypts Aine's message with Brianna's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 \ mod \ n$) to Cathleen encrypted with Cathleen's public key.
- Cathleen decrypts Brianna's message with Cathleen's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 \ mod \ n$) to Deirdre encrypted with Deirdre's public key.
- Deirdre decrypts Cathleen's message with Deirdre's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 + s_4 \ mod \ n$) to Aine encrypted with Aine's public key.
- Aine decrypts Deirdre's message with Aine's private key. She substracts (modulo n) the random number r added in the first step, obtaining in this way $s_1 + s_2 + s_3 + s_4$ (this will be in $[0, n]$).

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Aine adds a secret random number, say r (uniformly chosen in $[0, n]$) to her salary and sends it to Brianna encrypted with Brianna public key. Addition is modulo n. In this way, the outcome of $r + s_1 \ mod \ n$ will be a number uniformly distributed in $[0, n]$ and so Brianna will learn nothing about the actual value of s_1.
- Brianna decrypts Aine's message with Brianna's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 \ mod \ n$) to Cathleen encrypted with Cathleen's public key.
- Cathleen decrypts Brianna's message with Cathleen's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 \ mod \ n$) to Deirdre encrypted with Deirdre's public key.
- Deirdre decrypts Cathleen's message with Deirdre's private key, adds her salary (modulo n) and sends the result (i.e., $r + s_1 + s_2 + s_3 + s_4 \ mod \ n$) to Aine encrypted with Aine's public key.
- Aine decrypts Deirdre's message with Aine's private key. She substracts (modulo n) the random number r added in the first step, obtaining in this way $s_1 + s_2 + s_3 + s_4$ (this will be in $[0, n]$).
- Aine announces the result to the participants.

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- This protocol assumes that all of the participants are honest
- A participant can lie about her salary.
- Aine can announce a wrong addition.
- Participants can collude. E.g.,
  - Brianna and Deirdree can share their figures to find the salary of Cathleen

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

- Solving collusion.
  - Each salary is divided into shares.
  - The sum of each share is computed individually.
  - Different paths are used for different shares in a way that neighbors are different.

    To compute any s_i all neighbors of all paths are required.
  - Different number of shares imply different minimum coalition sizes for violating security

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed. Sum

Important observation

- This method is compliant with the privacy model selected: Secure multiparty computation
- This method is not compliant with other privacy models: differential privacy

We can define appropriate methods that satisfy multiple privacy models

- E.g., method that computes differentially private secure sum

Secure multiparty computation

Computation-driven approaches/multiple databases/distributed.

- Dining Cryptographers Problem.
  - (Chaum, 1985) Three cryptographers are sitting down to dinner at their favorite three-star restaurant. Their waiter informs them that arrangements have been made with the maître d'hôtel for the bill to be paid anonymously. One of the cryptographers might be paying the dinner, or it might have been NSA (U.S. National Security Agency). The three cryptographers respect each other's right to make an anonymous payment, but they wonder if NSA is paying.
- This problem (and previous ones) can be seen from a user's privacy perspective (more particularly, about protecting the data of the user). I.e., the cryptographers does not want to share whether they paid or not.

Secure multiparty computation

Computation-driven approaches/multiple databases: distributed.

- Machine learning and data mining methods.
- Parties can be seen as sharing the schema of a database.
- Two types of problems usually considered.
  - Vertically partitioned data. Parties (data holders) have information on the same individuals but different attributes.
  - Horizontally partitioned data. Parties (data holders) have information on different individuals but on the same attributes (i.e., the share the database schema).

Secure multiparty computation

Computation-driven approaches/multiple databases: distributed

Privacy leakage for the distributed approach is usually analyzed considering two types of adversaries.

Secure multiparty computation

Computation-driven approaches/multiple databases: distributed

Privacy leakage for the distributed approach is usually analyzed considering two types of adversaries.

- **Semi-honest adversaries.** Data owners follow the cryptographic protocol but they analyse all the information they get during its execution to discover as much information as they can.
- **Malicious adversaries.** Data owners try to fool the protocol (e.g. aborting it or sending incorrect messages on purpose) so that they can infer confidential information.

Data privacy mechanisms
Result-driven
Result-driven for association rule mining

Data Privacy

Respondent and owner privacy

- Data-driven or general-purpose
- Computation-driven or specific-purpose
- Result-driven (Ch. 3.5)

Data Privacy

Result-driven

- Prevent data mining procedures infer some knowledge that is valuable for the database owner
- Other uses: avoid discriminatory knowledge inferred from databases

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Data Privacy

Result-driven

- **Formalization.** Database $\mathcal{D}$, A data mining algorithm, with parameters Θ is said to have ability to derive knowledge K from $\mathcal{D}$ if and only if K is obtained from the output of the algorithm. Notation: $(A, \mathcal{D}, \Theta) \vdash K$.
- Any knowledge K such that $(A, \mathcal{D}, \Theta) \vdash K$ is in $KSet_{\mathcal{D}}$.

Definition. $\mathcal{D}$ a database, $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive knowledge to be hidden. The problem of hiding knowledge $\mathcal{K}$ from $\mathcal{D}$ consists on transforming $\mathcal{D}$ into a database $\mathcal{D}'$ such that

1. $\mathcal{K} \cap KSet_{\mathcal{D}'} = \emptyset$
2. the information loss from $\mathcal{D}$ to $\mathcal{D}'$ is minimal

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Data Privacy

Result-driven for association rules mining: Association rule hiding

- Recall that rules are mined when

$$Support(R) \geq thr - s$$

and

$$Confidence(R) \geq thr - c$$

for certain thresholds $thr - s$ and $thr - c$.

Two approaches:

- To reduce the support of the rule.
- To reduce the confidence of the rule.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.

Data Privacy

Result-driven for association rules mining: example

- **A formalization.** $\mathcal{D}$ a database; $thr - s$ threshold. Let $\mathcal{K} = \{K_1, \ldots, K_n\}$ sensitive itemsets, $\mathcal{A}$ non-sensitive itemsets.
- Transform $\mathcal{D} \rightarrow \mathcal{D}'$ such that
  1. $Support_{\mathcal{D}'}(K) < thr - s$ for all $K_i \in \mathcal{K}$
  2. The number of itemsets K in $\mathcal{A}$ such that $Support_{\mathcal{D}'}(K) < thr - s$ is minimized.

This problem is NP-hard (Atallah et al., 1999)

Because of this: heuristic approaches

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects
  the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Algorithm.**

  While HI is not hidden **do**
  HI' = HI;
  While $|HI'| > 2$ **do**
  P = subsets of HI with cardinality $|HI'| - 1$;
  HI'= $\arg\max_{hi \in P} Support(hi)$;
  Ts = transaction in T supporting HI that affects the mininum number of itemsets of cardinality 2;
  Set HI' = 0 in Ts;
  Propagate results forward;

- The algorithm does not cause false positives,
- only false negatives (rules no longer inferred)

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$
    $\rightarrow$ We select $HI' = \{a, c\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.

Transaction number	Items
T1	a, b, c, d
T2	a, b, c
T3	a, c, d

  - Subsets of HI with cardinality $|HI| - 1$: $\{a, b\}$, $\{b, c\}$, $\{a, c\}$.
  - $Support(\{a, b\}) = Support(\{b, c\}) = 2$, and $Support(\{a, c\}) = 3$ $\rightarrow$ We select $HI' = \{a, c\}$.
  - Set of transactions in T that support HI (and HI'): $\{T1, T2\}$.
  - Ts transaction in $\{T1, T2\}$ that affects the minimum number of itemsets of cardinality 2: $T2$ affects less itemsets than $T1$.

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:

Data Privacy

Result-driven for association rules mining: heuristic algorithm

- **Example.** Computation of the algorithm to hide $HI = \{a, b, c\}$.
- Remove one of the items in $HI' = \{a, c\}$ that are in $T2$:
  Both have the same support, we select one of them at random.
- Propagate the results forward: recompute supports

Data privacy mechanisms
Result-driven
Tabular data (Ch. 3.6)

Tabular data

- Aggregates of data with respect to a few variables.
  - Aggregates of data can lead to disclosure

Tabular data

- Aggregates of data with respect to a few variables. Ex. (Castro, 2012)

	P_1	P_2	P_3	P_4	P_5	Total
M_1	2	15	30	20	10	77
M_2	72	20	1	30	10	133
M_3	38	38	15	40	5	136
TOTAL	112	73	46	90	25	346

Cell (M_2, P_3): number of people with profession P_3 living in municipality M_2.

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	22	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

Cell (M_2, P_3): total salary received by people with profession P_3 living in M_2.

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **Internal attack.** A person whose data is in the database is able to use the information of the tables to infer some sensitive information about other individuals. A doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$

Tabular data

- Aggregates of data do not avoid disclosure
  - **External attack.** Combining the information of the two tables the adversary is able to infer some sensitive information.
    $\Rightarrow (M_2, P_3)$
  - **Internal attack.** A person whose data is in the database is able to use the information of the tables to infer some sensitive information about other individuals. A doctor infers the salary of another doctor.
    $\Rightarrow (M_1, P_1)$
  - **Internal attack with dominance.** This is an internal attack where a contribution of one person, say p_0, in a cell is so high that permits p_0 to obtain accurate bounds of the contribution of the others.
    $\Rightarrow (M_3, P_5)$ with 5 people. $salary(p_0) = 350$, then the salary of the other four is at most $363 - 350 = 13$.

Tabular data

- Privacy model / disclosure risk measure
- Data protection mechanism
- Information loss

Tabular data: privacy model

- **Rule (n,k)-dominance.** A cell is sensitive when n contributions represent more than the k fraction of the total. That is, the cell is sentitive when

$$\frac{\sum_{i=1}^{n} c_{\sigma(i)}}{\sum_{i=1}^{t} c_i} > k$$

where $\{\sigma(1),...,\sigma(t)\}$ is a permutation of $\{1,...,t\}$ such that $c_{\sigma(i-1)} \geq c_{\sigma(i)}$ for all $i = \{2,...,t\}$ (i.e., $c_{\sigma(i)}$ is the ith largest element in the collection $c_1,...,c_t$).

This rule is used with $n = 1$ or $n = 2$ and $k > 0.6$.

Tabular data: privacy model

- **Rule pq.** This rule is also known as the prior/posterior rule. It is based on two positive parameters p and q with $p < q$. Prior to the publication of the table, any intruder can estimate the contribution of contributors within the q percent. Then, a cell is considered sensitive if an intruder on the light of the released table can estimate the contribution of a contributor within p percent.

- **Rule p%.** This rule can be seen as a special case of the previous rule when no prior knowledge is assumed on any cell. Because of that, it can be seen as equivalent to the previous rule with $q = 100$.

Tabular data: data protection mechanism

- Protection of a tabular data
  - Perturbative. values are modified
    - ⋆ Post-tabular. Noise added after table preparation
      - Rounding
      - Controlled tabular adjustment (CTA). Replacing a table by another that is *similar*
    - ⋆ Pre-tabular. Noise added before table preparation
  - Non-perturbative. cell suppression

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression
- Primary suppression not enough:

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450	720	400	360	2290
M_2	1440	540	~~22~~	570	320	2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

- Secondary suppressions required:

	P_1	P_2	P_3	P_4	P_5	Total
M_1	360	450		400		2290
M_2	1440	540		570		2892
M_3	722	1178	375	800	363	3438
TOTAL	2522	2168	1117	1770	1043	8620

- Solutions built using optimization

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression
  - Decide which cells to suppress
  - Given a set of sensitive cells
  - Estimated values for suppressed cells should be outside a given interval
    (upper and lower protection levels;
    estimation based on non suppressed values + linear relationships)

  $\Rightarrow$ Problem formulated as an optimization problem

Tabular data: data protection mechanism

- Protection of a tabular data: cell suppression

$$\min \quad \sum_{i=1}^{n} w_i y_i$$

subject to

$$Ad^l = 0$$
$$(klo_i - a_i)y_i \leq d^{l,i} \leq (kup_i - a_i)y_i \quad \text{for all } i = 1, \ldots, n$$
$$d^{l,p} \leq -lo_p \quad \text{for all } p \in \mathcal{P}$$

$$Ad^u = 0$$
$$(klo_i - a_i)y_i \leq d^{u,i} \leq (kup_i - a_i)y_i \quad \text{for all } i = 1, \ldots, n$$
$$d^{u,p} \geq up_p \quad \text{for all } p \in \mathcal{P}$$

$$y_i \in \{0, 1\} \quad \text{for } i = 1, \ldots, n$$

Tabular data: information loss

- Minimal number of suppressions
- Weights associated to cells: *minimal weight* of suppressed cells

Summary

Terminology

- Main concepts
  - Naive anonymization does not work
  - Transparency and Privacy by design
  - (large number of) Privacy models
- Data privacy mechanisms
  - Data-driven (unknown use):
    - ⋆ databases (masking methods, IL, DR)
    - ⋆ tabular data (risk cells, IL)
  - Computation-driven (known use):
    - ⋆ differential privacy
    - ⋆ secure multiparty computation
  - Result-driven

References

References

- V. Torra (2017) Data privacy: Foundations, New Developments and the Big Data Challenge, Springer.

- V. Torra, G. Navarro-Arribas (2016) Big Data Privacy and Anonymization, Privacy and Identity Management 15-26
https://doi.org/10.1007/978-3-319-55783-0_2

- V. Torra, G. Navarro-Arribas, K. Stokes (2018) Data Privacy, in A. Saida, V. Torra (eds) Data Science in Practice, Springer.
https://link.springer.com/chapter/10.1007/978-3-319-97556-6_7

Thank you

http://www.ppdm.cat/dp/